TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

---------- ----------

BÁO CÁO ĐỒ ÁN THỰC HÀNH

ĐỒ ÁN THỰC HÀNH #1

Môn: Phân tích dữ liệu ứng dụng

05/10/2024 – *21/11/2024*

MÃ HỌC PHẦN: CSC12110
Nhóm: 3
Thành viên:

ID	Họ tên
18127008	Lê Mạnh Hoàng
21127211	Nguyễn Vũ Tường An
21127699	Lô Thủy Tiên

Giảng viên:
Vũ Thị Mỹ Hằng
Hồ Thị Hoàng Vy

Báo cáo:
Lô Thuỷ Tiên

MỤC LỤC

THÔNG TIN VỀ ĐỒ ÁN

Mã học phần: CSC12110

Tên học phần: PHÂN TÍCH DỮ LIỆU ỨNG DỤNG

Chủ đề: ĐỒ ÁN THỰC HÀNH #1

Hình thức:

Nộp file .ipynb và link Colab có quyền chỉnh sửa, đánh giá tỷ lệ tham gia của từng thành viên và đưa ra nhận xét, kết luận đầy đủ.

Mô tả:

1. **Khám phá và tiền xử lý dữ liệu:** Thực hiện EDA để kiểm tra và xử lý dữ liệu missing, trùng lặp, và outliers. Kiểm tra sự cân bằng giữa nhãn click/không click quảng cáo.
2. **Phân tích EDA (Exploratory Data Analysis):** Thực hiện phân tích đơn biến và hai biến; loại bỏ thuộc tính không cần thiết và phân tích tỷ lệ click quảng cáo theo các yếu tố như tuổi, thu nhập, giới tính.
3. **Quan sát hành vi người dùng:** Phân tích thời gian sử dụng website theo tuổi, thu nhập, vị trí địa lý và thời điểm trong ngày. Nhận xét về các chủ đề quảng cáo được quan tâm nhiều nhất.
4. **Xây dựng mô hình dự đoán:** Cài đặt ít nhất hai mô hình dự đoán khả năng click quảng cáo của người dùng để so sánh. Đánh giá ảnh hưởng của thuộc tính thu nhập đến khả năng dự đoán và xác định các thuộc tính quan trọng.
5. **Đánh giá mô hình:** Đánh giá chất lượng mô hình bằng cross-validation với các độ đo precision, recall, f1 trên tập train và test. Chọn mô hình tối ưu và nêu kết luận.

Giảng viên phụ trách: Cô Hồ Thị Hoàng Vy, Cô Vũ Thị Mỹ Hằng

THÔNG TIN NHÓM

Nhóm: 3

MSSV	Họ tên	Email	Ghi chú
18127008	Lê Mạnh Hoàng	lmhoang18@clc.fitus.edu.vn	
21127211	Nguyễn Vũ Tường An	nvtan21@clc.fitus.edu.vn	
21127699	**Lô Thủy Tiên**	**lttien21@clc.fitus.edu.vn**	

GIỚI THIỆU TỔNG QUÁT : ĐỒ ÁN THỰC HÀNH #1

I. Giới thiệu đồ án:

1. *Mục tiêu của báo cáo:*

Báo cáo tập trung vào việc dự đoán khả năng người dùng sẽ click vào quảng cáo thông qua phân tích dữ liệu và xây dựng mô hình dự đoán. Mục tiêu cụ thể bao gồm:

- **Phân tích dữ liệu:** Khám phá và phân tích các yếu tố ảnh hưởng đến hành vi click quảng cáo của người dùng dựa trên các thuộc tính trong dataset.
- **Xây dựng mô hình dự đoán:** Từ dữ liệu đã qua tiền xử lý, xây dựng một mô hình dự đoán có độ chính xác cao, giúp tối ưu hóa chiến dịch quảng cáo bằng cách nhận diện người dùng có khả năng click cao.
- **Đánh giá và tối ưu hóa:** Đánh giá hiệu quả của mô hình dự đoán và tối ưu hóa để tăng độ chính xác, từ đó giúp dự đoán tốt hơn hành vi người dùng đối với quảng cáo.

2. *Giới thiệu dataset:*

Dataset ***"2425_QC.csv"*** được cung cấp chứa các thông tin sau về người dùng:

- **Daily Time Spent on Site:** Thời gian trung bình mà người dùng dành trên website mỗi ngày.
- **Age:** Tuổi của người dùng.
- **Area Income:** Thu nhập trung bình của khu vực nơi người dùng sinh sống.
- **Daily Internet Usage:** Thời gian trung bình mà người dùng dành trên internet mỗi ngày.
- **Ad Topic Line:** Chủ đề của quảng cáo mà người dùng nhìn thấy.
- **City:** Thành phố nơi người dùng đang sinh sống.
- **Male (0,1):** Giới tính của người dùng (1 là nam, 0 là nữ).

- **Country:** Quốc gia nơi người dùng đang sinh sống.
- **Timestamp:** Thời điểm người dùng nhìn thấy quảng cáo.
- **Clicked on Ad:** Biến mục tiêu (label), cho biết người dùng có click vào quảng cáo hay không (1 là click, 0 là không).

	A	B	C	D	E	F	G	H	I	J
1	Daily Time	Age	Area Incor	Daily Inter	Ad Topic Lir	City	Male	Country	Timestamp	Clicked on Ad
2	68.95	35	61833.9	256.09	Cloned 5thg	Wrightbur	0	Tunisia	27/3/2016 0:53	0
3	80.23	31	68441.85	193.77	Monitored	West Jodi	1	Nauru	4/4/2016 1:39	0
4	69.47	26	59785.94	236.5	Organic bot	Davidton	0	San Marin	13/3/2016 20:35	0
5	74.15	29	54806.18	245.89	Triple-buffe	West Terr	1	Italy	10/1/2016 2:31	0
6	68.37	35	73889.99	225.58	Robust logi	South Mar	0	Iceland	3/6/2016 3:36	0
7	59.99	23	59761.56	226.74	Sharable cli	Jamieberg	1	Norway	19/5/2016 14:30	0
8	88.91	33	53852.85	208.36	Enhanced d	Brandonst	0	Myanmar	28/1/2016 20:59	0
9	66	48	24593.33	131.76	Reactive loc	Port Jeffe	1	Australia	7/3/2016 1:40	1
10	74.53	30	68862	221.51	Configurabl	West Coli	1	Grenada	18/4/2016 9:33	0
11	69.88	20	55642.32	183.82	Mandatory	Ramirezto	1	Ghana	11/7/2016 1:42	0

3. *Phương pháp thực hiện:*

Quá trình thực hiện báo cáo bao gồm các bước chính sau đây:

- **Chuẩn bị dữ liệu (Data Preparation):** Tải dữ liệu từ nguồn cung cấp, kiểm tra cấu trúc dữ liệu (số lượng hàng, cột), và xác định các thuộc tính cần cho quá trình phân tích và dự đoán.
- **Khám phá và tiền xử lý dữ liệu (Data Preprocessing):** Thực hiện phân tích sơ bộ dữ liệu, kiểm tra dữ liệu thiếu, trùng lặp, ngoại lai và xử lý các vấn đề này để làm sạch dữ liệu.
- **Phân tích EDA (Exploratory Data Analysis):** Thực hiện phân tích đơn biến và hai biến để hiểu sâu hơn về các thuộc tính quan trọng và mối quan hệ của chúng với biến mục tiêu.
- **Xây dựng mô hình dự đoán:** Chọn và huấn luyện các mô hình học máy để dự đoán khả năng người dùng sẽ click vào quảng cáo.
- **Đánh giá mô hình:** Sử dụng các chỉ số đo lường như precision, recall, và F1-score để đánh giá hiệu quả của mô hình và chọn mô hình tối ưu.

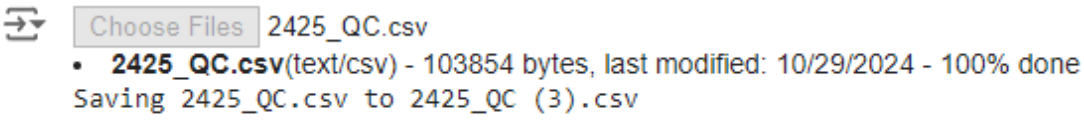
BÁO CÁO THỰC HIỆN ĐỒ ÁN

I. Chuẩn bị dữ liệu (Data Preparation):

- ***Tải dữ liệu từ nguồn cung cấp :*** Đọc tập dữ liệu từ file (ví dụ: CSV) bằng các thư viện pandas.

```
[17] #Uploading files from your local file system
     from google.colab import files
     uploaded = files.upload()

     import pandas as pd
     data = pd.read_csv("2425_QC.csv", index_col=0)
```



- ***Kiểm tra cấu trúc dữ liệu:*** Hiểu rõ dữ liệu có bao nhiêu hàng, cột và các thuộc tính chính, xác định các thuộc tính quan trọng phục vụ cho quá trình phân tích và dự đoán.

```
print("Number of rows and columns in the data:", data.shape)

print("Information about the data structure:")
data.info()

print("Names of the attributes in the dataset:", data.columns)

print("Some top rows of the data:")
data.head()
```

- Số hàng và số cột của dữ liệu: (1002, 9)
- Thông tin về cấu trúc dữ liệu:
  <class 'pandas.core.frame.DataFrame'>
  Index: 1002 entries, 68.95 to 36.91

```
Data columns (total 9 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Age                   1001 non-null   float64
 1   Area Income           1002 non-null   float64
 2   Daily Internet Usage  1002 non-null   float64
 3   Ad Topic Line         1002 non-null   object
 4   City                  1002 non-null   object
 5   Male                  1002 non-null   int64
 6   Country               1002 non-null   object
 7   Timestamp             1002 non-null   object
 8   Clicked on Ad         1002 non-null   int64
dtypes: float64(3), int64(2), object(4)
memory usage: 78.3+ KB
```

- Tên các thuộc tính trong dataset: Index(['Age', 'Area Income', 'Daily Internet Usage', 'Ad Topic Line', 'City', 'Male', 'Country', 'Timestamp', 'Clicked on Ad'],
- Một số hàng đầu của dữ liệu:

Một số hàng đầu của dữ liệu:

Daily Time Spent on Site	Age	Area Income	Daily Internet Usage	Ad Topic Line	City	Male	Country	Timestamp	Clicked on Ad
68.95	35.0	61833.90	256.09	Cloned 5thgeneration orchestration	Wrightburgh	0	Tunisia	27/3/2016 0:53	0
80.23	31.0	68441.85	193.77	Monitored national standardization	West Jodi	1	Nauru	4/4/2016 1:39	0
69.47	26.0	59785.94	236.50	Organic bottom-line service-desk	Davidton	0	San Marino	13/3/2016 20:35	0
74.15	29.0	54806.18	245.89	Triple-buffered reciprocal time-frame	West Terrifurt	1	Italy	10/1/2016 2:31	0
68.37	35.0	73889.99	225.58	Robust logistical utilization	South Manuel	0	Iceland	3/6/2016 3:36	0

II. Khám phá và tiền xử lý dữ liệu (Data Preprocessing):

1. Handling Missing Value & Empty Data:

- Mô tả số lượng missing value của mỗi thuộc tính:

```
# Kiểm tra các giá trị bị thiếu
print("\nMissing Values:")
missing_values = data.isnull().sum()
print(missing_values)
```

```
Missing Values:
Age                       1
Area Income               0
Daily Internet Usage      0
Ad Topic Line             0
City                      0
Male                      0
Country                   0
Timestamp                 0
Clicked on Ad             0
dtype: int64
```

=> Nhận xét: Chỉ có 1 giá trị bị thiếu trong cột Age.

- Xử lý dữ liệu missing : Việc điền giá trị thiếu trong cột "Age" bằng giá trị trung vị giúp đảm bảo rằng dữ liệu không bị thiên lệch do giá trị ngoại vi và vẫn duy trì tính chính xác cho các phân tích và mô hình dự đoán tiếp theo.

2. ***Handling Duplications | Errors:***
   - Mô tả số lượng Duplications | Errors:

Mô tả số lượng Duplications | Errors

```
# Kiểm tra các dòng trùng lặp
duplicates = data.duplicated()
print("\nDuplicate rows:")
print(duplicates)
```

```
Duplicate rows:
Daily Time Spent on Site
68.95    False
80.23    False
69.47    False
74.15    False
68.37    False
         ...
51.30    False
51.63    False
55.55    False
45.01    False
36.91     True
Length: 1002, dtype: bool
```

=> Nhận xét: Dữ liệu cho thấy rằng dòng có giá trị "Daily Time Spent on Site" là 36.91 có bản sao trùng lặp, được đánh dấu là True, trong khi các dòng khác đều được đánh dấu là False, cho thấy chúng là các giá trị duy nhất.

- Xử lý dữ liệu Handling Duplications | Errors : Sử dụng phương thức drop_duplicates(keep='first') để giữ lại dòng đầu tiên của các dòng trùng lặp và xóa các dòng còn lại.

3. *Incorrect | Invalid Values - Validating Data - Normalizing Data*

- Incorrect | Invalid Values:

Các giá trị null trong dữ liệu:

```
Daily Time Spent on Site    0
Age                         1
Area Income                 0
Daily Internet Usage        0
Ad Topic Line               0
City                        0
Male                        0
Country                     0
Timestamp                   0
Clicked on Ad               0
dtype: int64
```

- Các giá trị null sau khi xử lí:

Các giá trị null sau khi xử lý:

```
Daily Time Spent on Site    0
Age                         0
Area Income                 0
Daily Internet Usage        0
Ad Topic Line               0
City                        0
Male                        0
Country                     0
Timestamp                   0
Clicked on Ad               0
dtype: int64
```

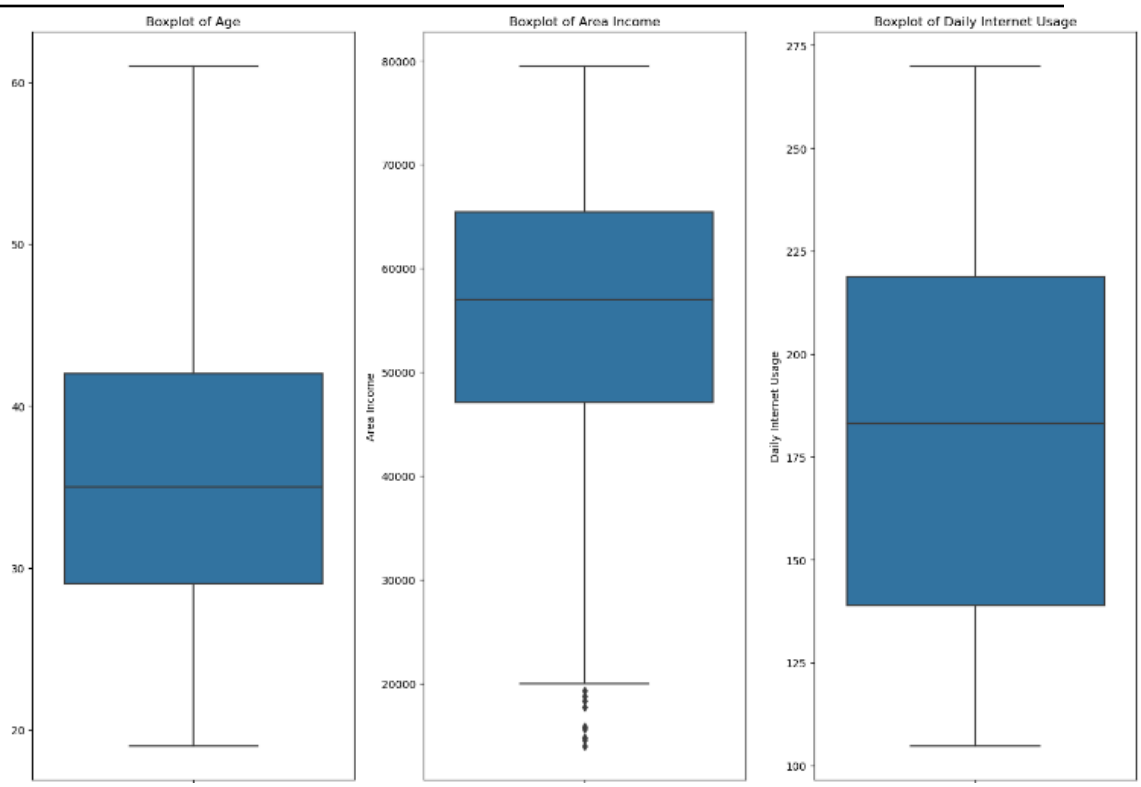
- Sau khi chuẩn hóa dữ liệu ta có:

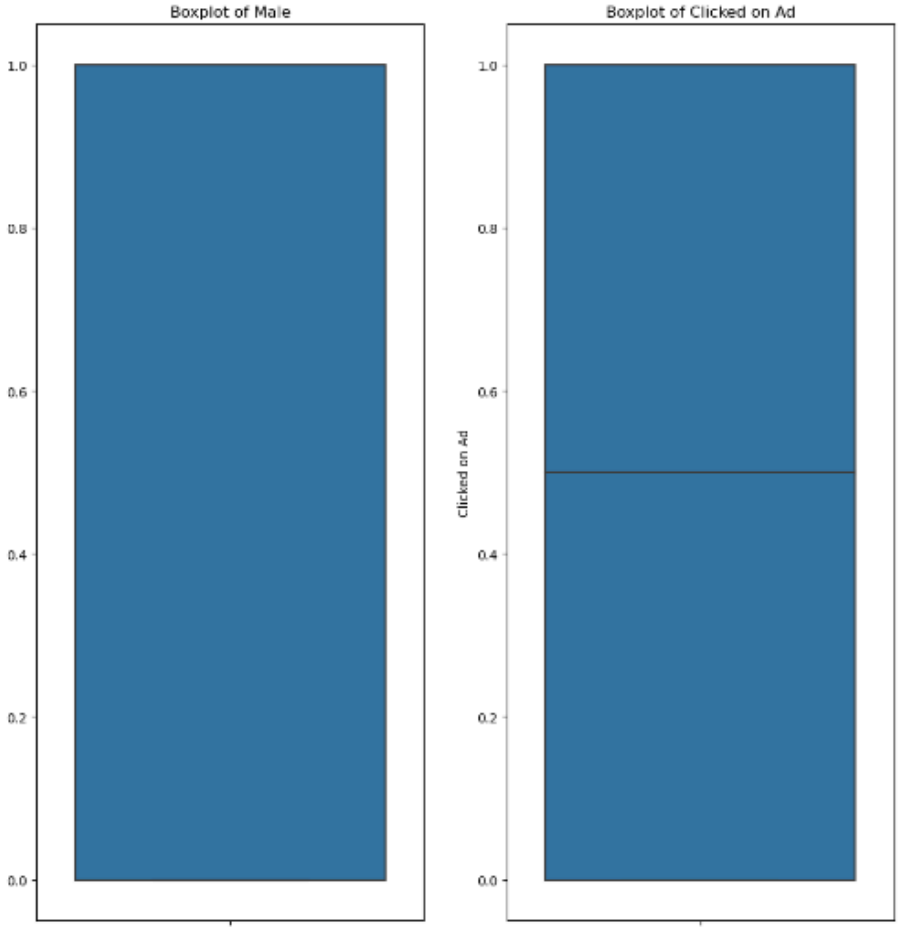
	Daily Time Spent on Site	Age	Area Income	Daily Internet Usage	Ad Topic Line	City	Male	Country	Timestamp	Clicked on Ad
0	0.249621	-0.116722	0.509802	1.732781	Cloned 5thgeneration orchestration	Wrightburgh	0	Tunisia	27/3/2016 0:53	0
1	0.960357	-0.572236	1.002880	0.312991	Monitored national standardization	West Jodi	1	Nauru	4/4/2016 1:39	0
2	0.282385	-1.141627	0.356985	1.286476	Organic bottom-line service-desk	Davidton	0	San Marino	13/3/2016 20:35	0
3	0.577265	-0.799992	-0.014600	1.500402	Triple-buffered reciprocal time-frame	West Terrifurt	1	Italy	10/1/2016 2:31	0
4	0.213076	-0.116722	1.409415	1.037694	Robust logistical utilization	South Manuel	0	Iceland	3/6/2016 3:36	0

4. Handling Outliers | Non Relevant Data:

Mô tả số lượng outlier value bằng 2 phương pháp:

- Trực quan hóa bằng boxplot


Boxplot of Age
Boxplot of Area Income
Boxplot of Daily Internet Usage
Area Income
Daily Internet Usage


Boxplot of Male
Boxplot of Clicked on Ad
Clicked on Ad

- Thống kê bằng Z-score hoặc IQR
  + Bằng Z-score

```
Outliers detected by Z-score method:
Age: 0 outliers
Area Income: 3 outliers
Daily Internet Usage: 0 outliers
Male: 0 outliers
Clicked on Ad: 0 outliers
```

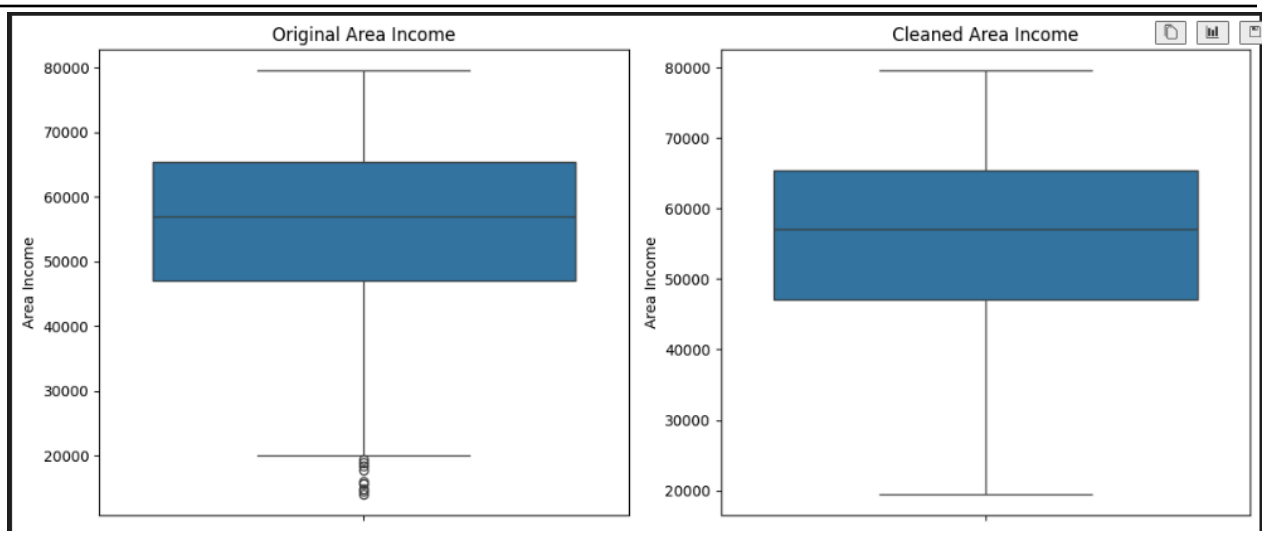
  + Bằng IQR

```
Outliers detected by IQR method:
Age: 0 outliers
Area Income: 9 outliers
Daily Internet Usage: 0 outliers
Male: 0 outliers
Clicked on Ad: 0 outliers
```

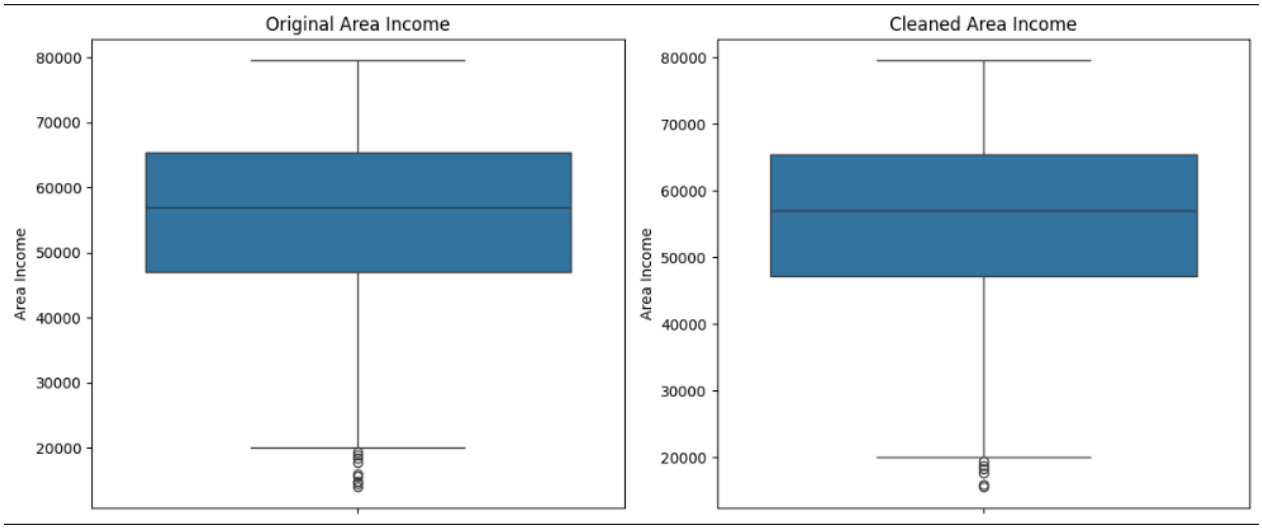
- Tiến hành xử lý dữ liệu Outliers bằng 2 phương pháp removing và capping ta được kết quả như sau:

```
Original shape: (1002, 9)
Shape if using Z-score method after removing outliers: (999, 9)
Shape if using Z-score method after capping outliers: (1002, 9)
Shape if using IQR method after removing outliers: (993, 9)
Shape if using IQR method after capping outliers: (1002, 9)
```

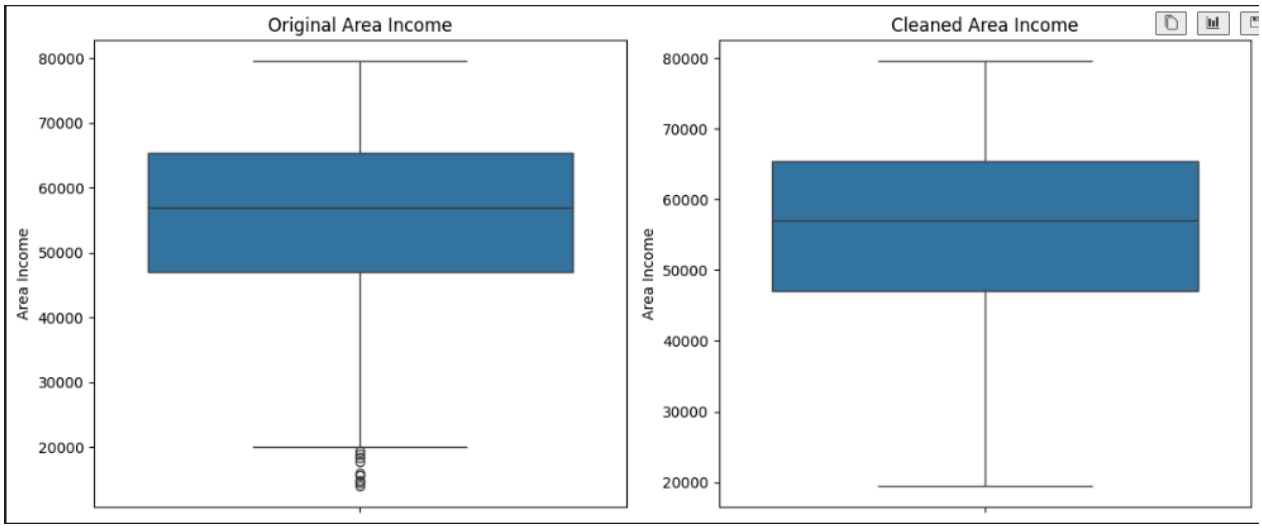
- Kết quả sau khi loại bỏ Outliers
  + Bằng phương pháp Z-score sau khi capping:



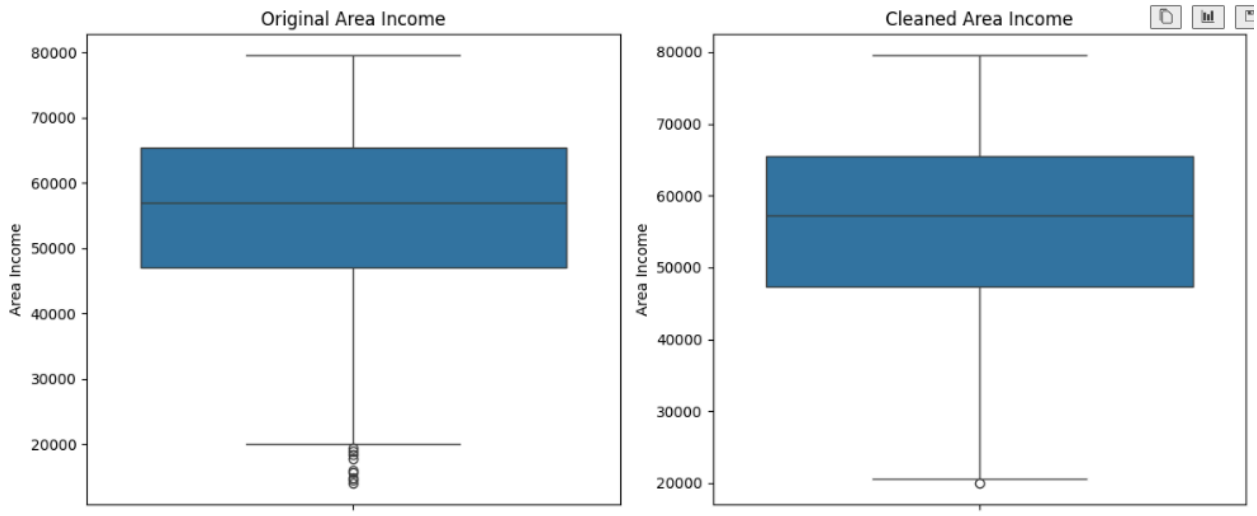
+ Bằng phương pháp Z-score sau khi removing:



+ Bằng phương pháp IQR sau khi capping:



+ Bằng phương pháp IQR sau khi removing:



⇒ Nhận xét: ta có thể thấy với cả 2 phương pháp loại bỏ Outliers bằng Z-score hoặc IQR, nếu chúng ta thực hiện capping thay vì removing thì số lượng giá trị Outliers sẽ được xử lý triệt để hơn.

5. *Incorrect Types:*

Tiến hành định dạng lại cột Timestamp với định dạng 'dd/mm/yyyy hh:mm'

- Cột Timestamp trước khi định dạng:

Male	Country	Timestamp	Clicked on Ad
0	Tunisia	27/3/2016 0:53	0
1	Nauru	4/4/2016 1:39	0
0	San Marino	13/3/2016 20:35	0
1	Italy	10/1/2016 2:31	0
0	Iceland	3/6/2016 3:36	0
1	Norway	19/5/2016 14:30	0
0	Myanmar	28/1/2016 20:59	0
1	Australia	7/3/2016 1:40	1
1	Grenada	18/4/2016 9:33	0
1	Ghana	11/7/2016 1:42	0

- Cột Timestamp sau khi định dạng:

Male	Country	Timestamp	Clicked on Ad
0	Tunisia	27/03/2016 00:53	0
1	Nauru	04/04/2016 01:39	0
0	San Marino	13/03/2016 20:35	0
1	Italy	10/01/2016 02:31	0
0	Iceland	03/06/2016 03:36	0
1	Norway	19/05/2016 14:30	0
0	Myanmar	28/01/2016 20:59	0
1	Australia	07/03/2016 01:40	1
1	Grenada	18/04/2016 09:33	0
1	Ghana	11/07/2016 01:42	0

6. ***Standardizing Data***

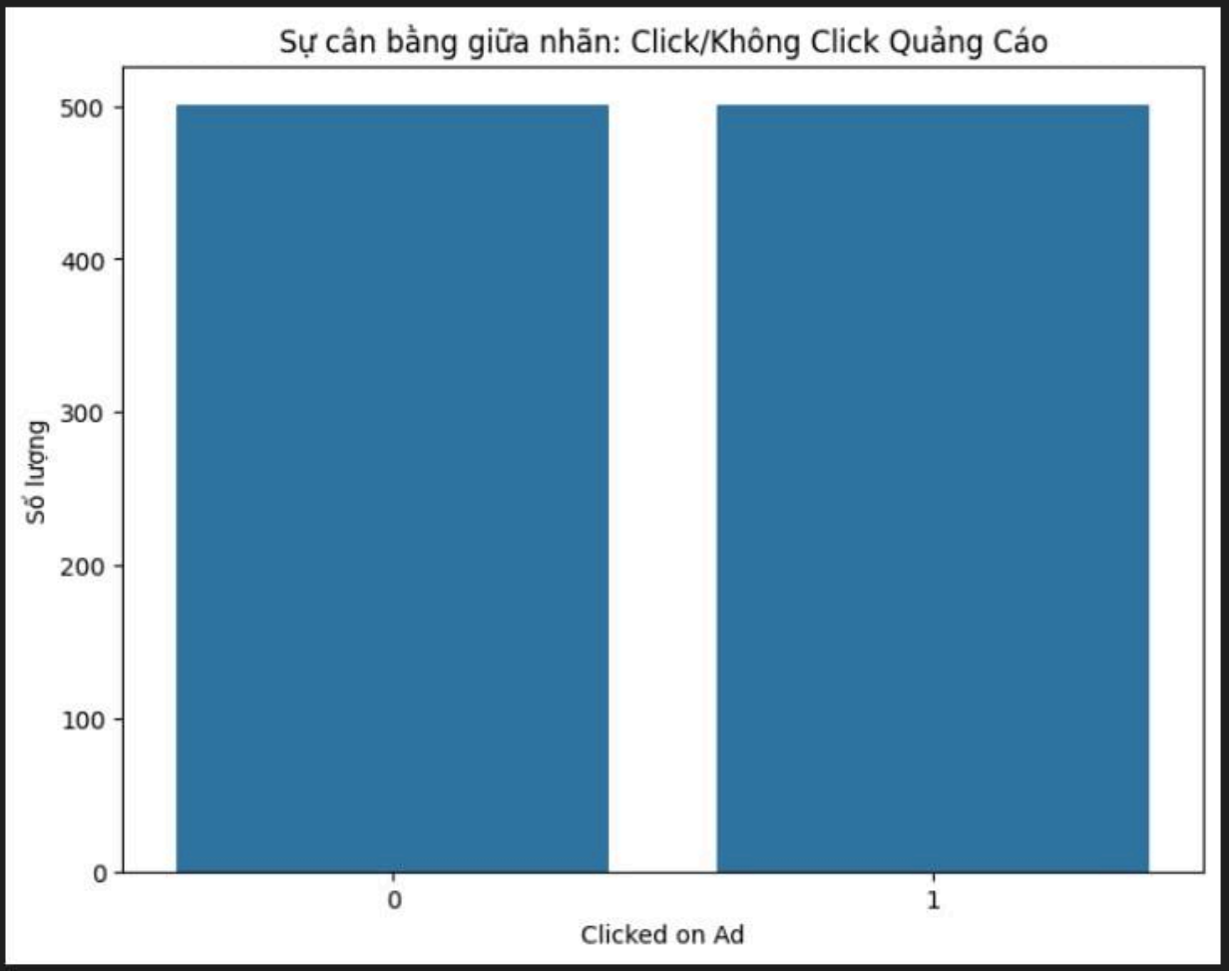
III. Phân tích EDA (Exploratory Data Analysis)

1. *Phân tích đơn biến (Univariate Analysis), Bivariate Analysis:*

a. *Kiểm tra sự cân bằng giữa nhãn: click/không click quảng cáo*

- Sử dụng value_counts() để đếm số lượng của từng nhãn (Clicked on Ad):

```
Số lượng của từng nhãn (Clicked on Ad):
Clicked on Ad
0    501
1    501
Name: count, dtype: int64
```



b. <u>Phân tích các biến số (age, income, gender, daily internet usage, etc.).</u>

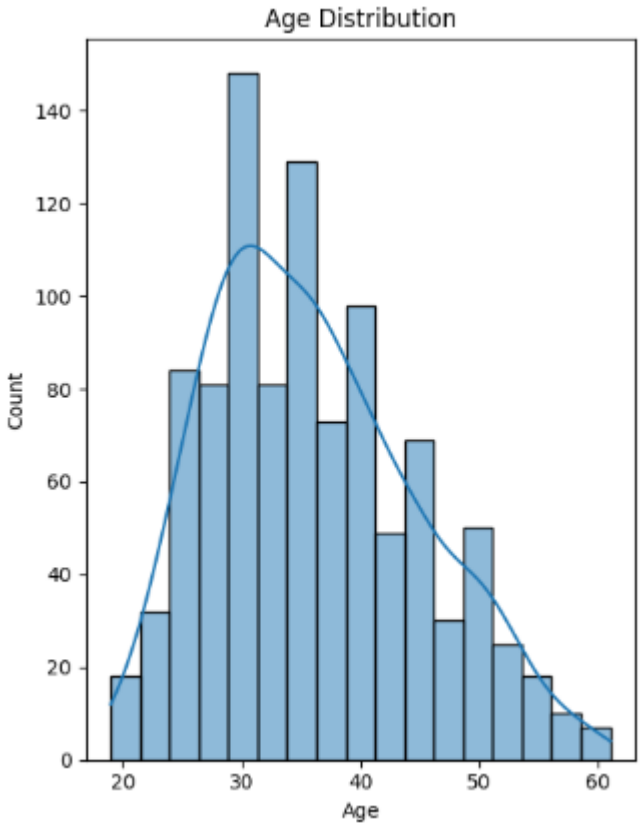
Gợi ý:

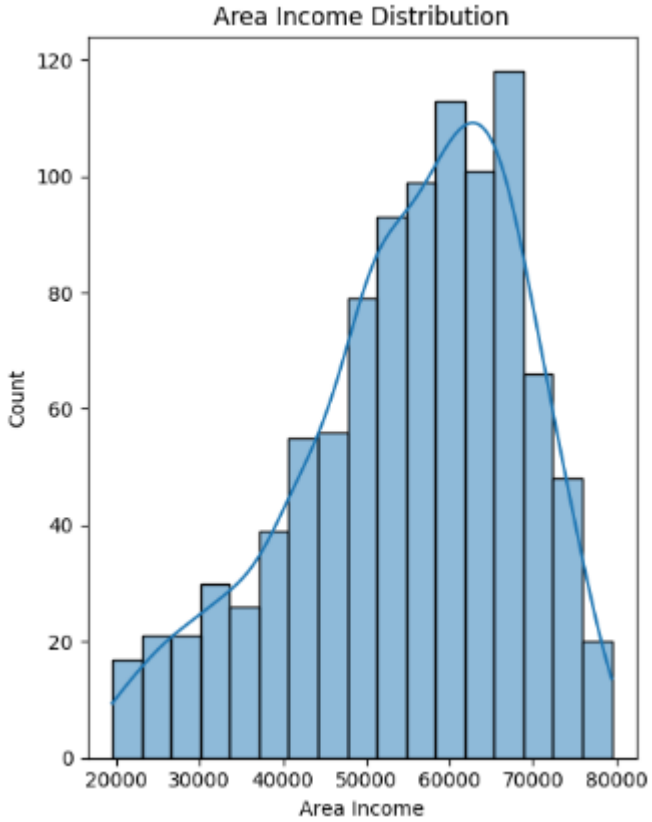
- Tính toán thống kê cơ bản: trung bình, độ lệch chuẩn, phân phối.
- Vẽ biểu đồ:
  - **Biểu đồ histogram** cho các biến số liên tục như độ tuổi (age), thu nhập (income), thời gian sử dụng (daily_time_spent).
  - **Biểu đồ bar plot** cho các biến phân loại như giới tính (Male).

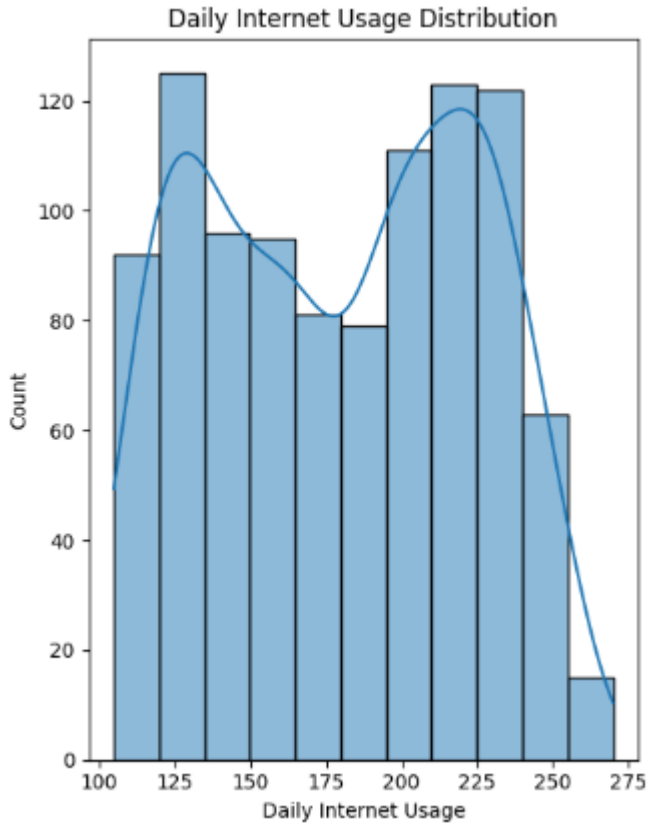
Kết quả thực thi:

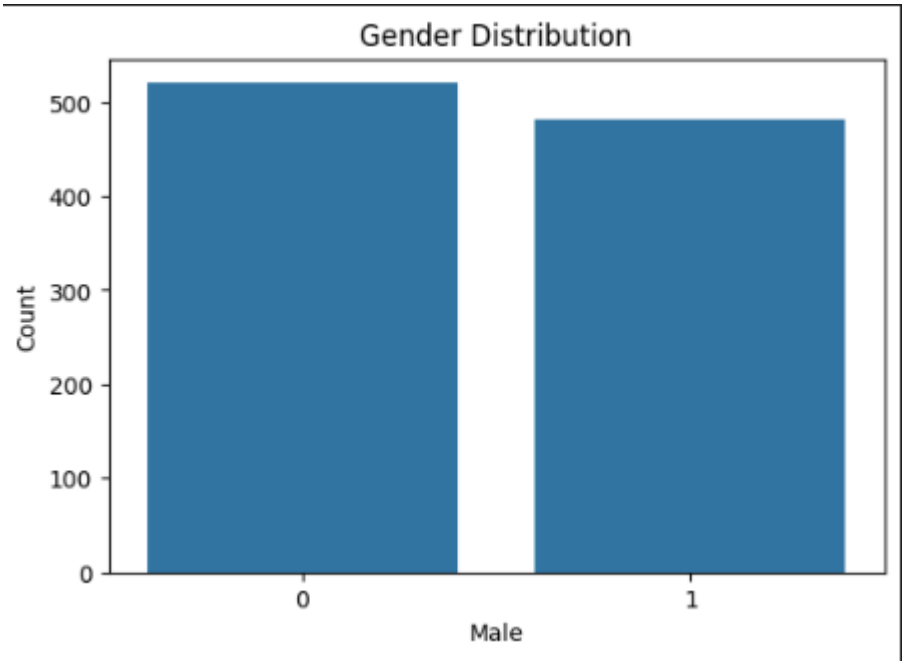
	Age	Area Income	Daily Internet Usage	Male
count	1002.000000	1002.000000	1002.000000	1002.000000
mean	36.023952	55028.288850	180.031637	0.481038
std	8.781362	13326.918589	43.893820	0.499890
min	19.000000	19504.987500	104.780000	0.000000
25%	29.000000	47073.067500	138.905000	0.000000
50%	35.000000	57012.300000	183.130000	0.000000
75%	42.000000	65451.787500	218.797500	1.000000
max	61.000000	79484.800000	269.960000	1.000000

	Clicked on Ad
count	1002.00000
mean	0.50000
std	0.50025
min	0.00000
25%	0.00000
50%	0.50000
75%	1.00000
max	1.00000


Age Distribution
Count
Age
0
20
40
60
80
100
120
140
20
30
40
50
60


Area Income Distribution
Count
Area Income
0
20
40
60
80
100
120
20000
30000
40000
50000
60000
70000
80000


Daily Internet Usage Distribution
Count
Daily Internet Usage
0
20
40
60
80
100
120
100
125
150
175
200
225
250
275


Gender Distribution
Count
Male
0
100
200
300
400
500
0
1

Nhận xét tổng quan:

- Trung bình độ tuổi từ 30 đến 40 tuổi là độ tuổi sử dụng Internet với các mục đích khác nhau nhiều nhất, và số lượng người dùng Internet giảm dần khi tuổi càng cao.
- Phần lớn những người truy cập Internet có thu nhập từ 50000 - 70000 USD/năm.
- Trung bình người dùng Internet truy cập từ 100 - 225 phút mỗi ngày.
- Chênh lệch về giới tính của người dùng Internet hầu như không quá lớn, có thể nói gần như là bằng nhau về số lượng nam và nữ.

Kết luận

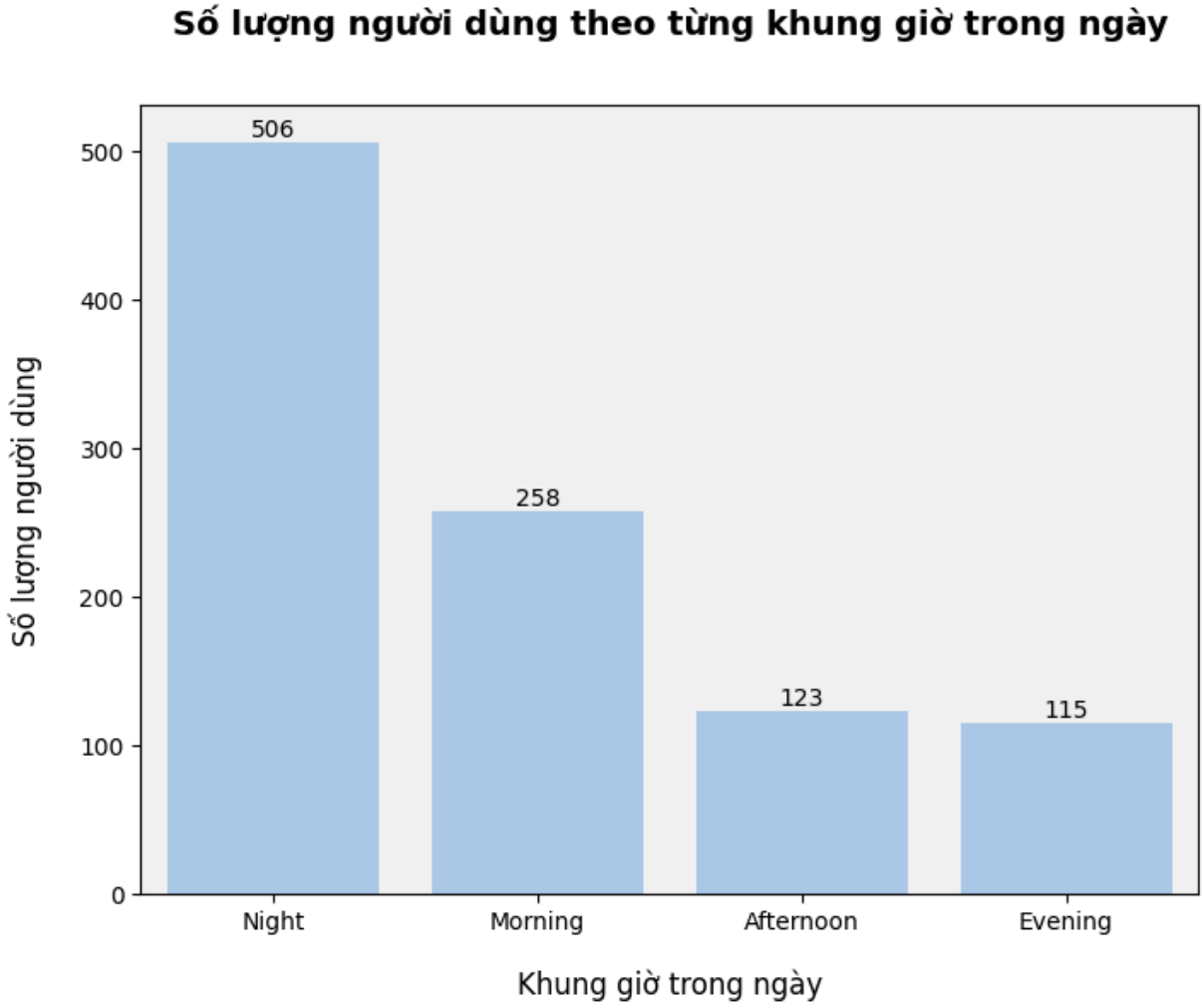
- Phần lớn người dùng truy cập Internet đều là những người đang trong độ tuổi làm việc và có thu nhập ổn định, do đó việc truy cập Internet phần lớn phục vụ cho công việc của họ.

c. ***Phân tích thời điểm sử dụng website trong ngày:***

Gợi ý:

- Phân chia thành các khung giờ (Sáng, Trưa, Chiều, Đêm).
- Vẽ bar plot hoặc count plot để quan sát phân bố số lượng người dùng.
- Xây dựng Hàm Phân loại Thời điểm: Hàm classify_time_of_day(timestamp) được định nghĩa để phân loại một thời điểm trong ngày vào một trong các khung giờ:
  + Sáng (Morning): Từ 6:00 AM đến 11:59 AM.
  + Trưa (Afternoon): Từ 12:00 PM đến 2:59 PM.
  + Chiều (Evening): Từ 3:00 PM đến 5:59 PM.
  + Đêm (Night): Từ 6:00 PM đến 5:59 AM ngày hôm sau.
- Đếm số lượng bản ghi theo mỗi khung giờ: Sử dụng value_counts(), đếm số lượng bản ghi (người dùng) trong mỗi khung giờ.

Kết quả thực thi:



Nhận xét tổng quan :

- Đêm là thời điểm phổ biến nhất để người dùng truy cập website, có thể do đây là thời gian rảnh rỗi nhất trong ngày.
- Sáng cũng là thời điểm khá phổ biến, có thể do thói quen truy cập internet vào đầu ngày.
- Trưa và Chiều là thời điểm ít người dùng truy cập website nhất, có thể do bận rộn với các hoạt động hàng ngày

Kết luận:

- Người dùng có xu hướng sử dụng website vào **ban đêm** nhiều hơn các thời gian khác trong ngày.
- Sự giảm dần số lượng người dùng từ "**Night**" đến "**Evening**" và "**Afternoon**" có thể phản ánh thói quen sử dụng internet của người dùng, khi họ có xu hướng truy

cập website vào thời gian tự do, như ban đêm sau khi hoàn thành công việc trong ngày.

d. Phân tích chủ đề quảng cáo. Xác định các chủ đề quảng cáo phổ biến.

Gợi ý:

- Tạo nhóm chủ đề từ Ad Topic Line.
- Dùng Word Cloud để xem các chủ đề quảng cáo.

Kết quả thực thi:

Nhận xét tổng quan:

- **Từ khóa nổi bật:** Các từ khóa như "coherent", “asynchronous”, “5thgeneration”, “system engine”, “hierarchy”, “Internet solution”, “secure”, “application”, xuất hiện nhiều lần và có kích thước lớn trong Word Cloud. Điều này cho thấy các chủ đề quảng cáo thường tập trung vào các khái niệm liên quan đến công nghệ, hệ thống, và giải pháp.
- **Xu hướng chủ đề:** Các chủ đề quảng cáo phổ biến thường liên quan đến việc cải thiện hiệu suất, tính linh hoạt, và khả năng tương tác của các hệ thống và phần mềm. Các từ khóa như "coherent", “stable”, "dynamic", “moderator”,

"capability", "support" cho thấy sự tập trung vào việc cung cấp các giải pháp công nghệ tiên tiến và hiệu quả.

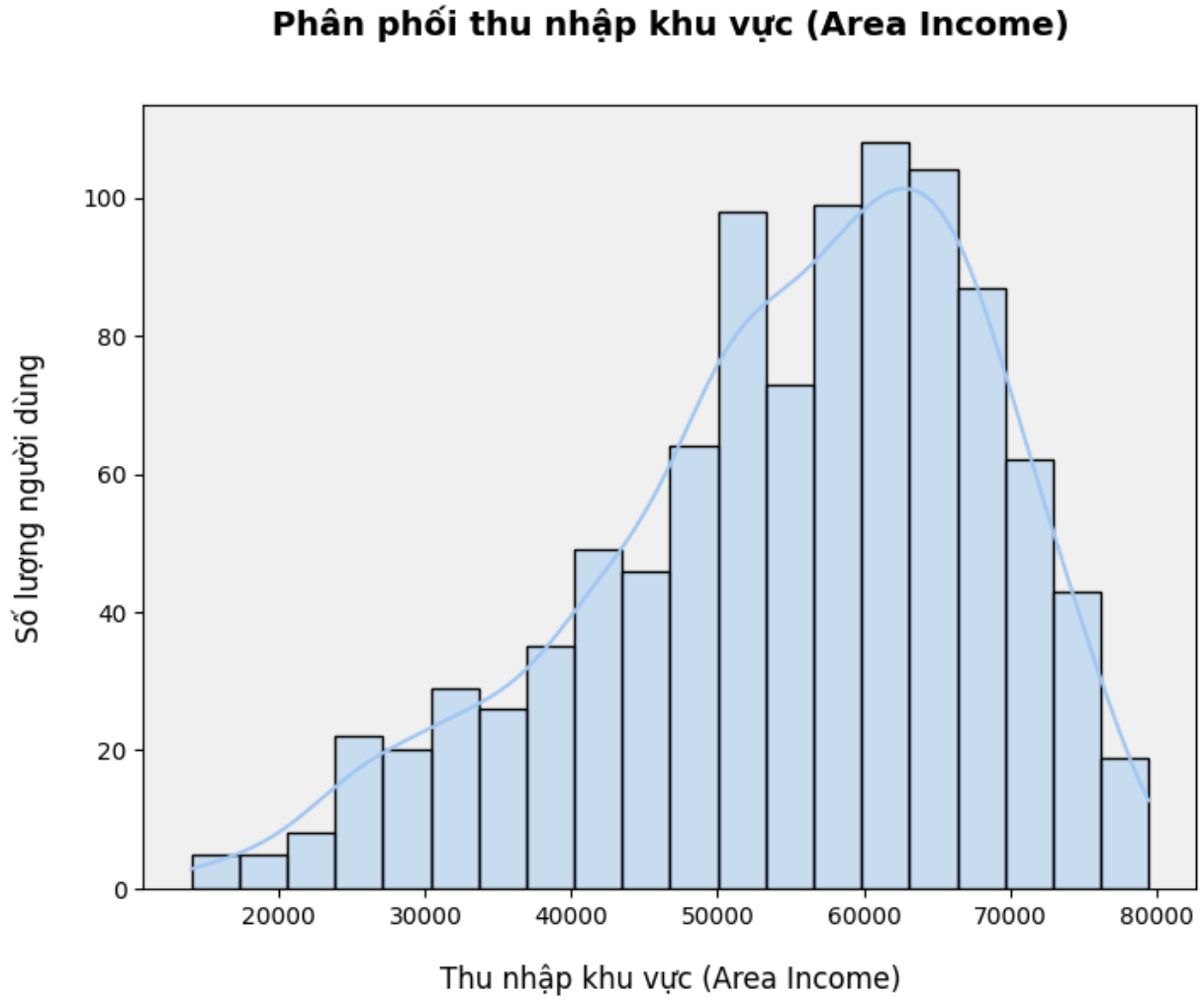
- **Đa dạng chủ đề:** Mặc dù có một số từ khóa xuất hiện nhiều lần, word cloud cũng cho thấy sự đa dạng trong các chủ đề quảng cáo. Các từ khóa như "orchestration", "multi-state", "success", "impactful" cho thấy các quảng cáo không chỉ tập trung vào một khía cạnh duy nhất mà còn bao gồm nhiều khía cạnh khác nhau của công nghệ và giải pháp.
- **Tầm quan trọng của từ khóa:** Kích thước của các từ khóa trong word cloud phản ánh tầm quan trọng và tần suất xuất hiện của chúng trong các dòng chủ đề quảng cáo. Các từ khóa lớn hơn như "coherent", "asynchronous", "5thgeneration", "success" cho thấy chúng là những yếu tố quan trọng và thường được nhấn mạnh trong các quảng cáo.

e. Phân tích thu nhập khu vực (Area Income). Đánh giá phân phối thu nhập trung bình của người dùng.

Gợi ý:

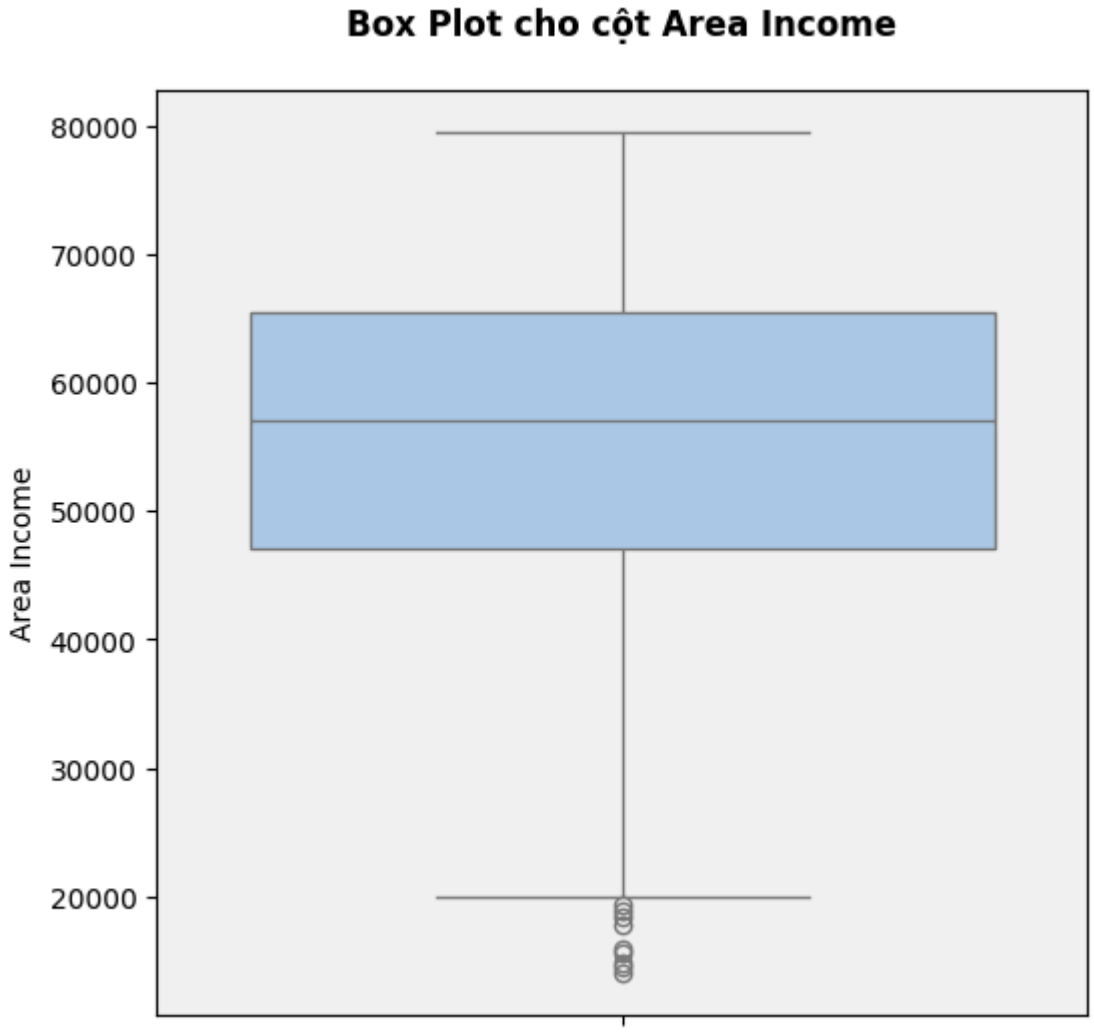
- Vẽ **histogram** hoặc **box plot** cho cột **Area Income**.
- Quan sát sự phân bố thu nhập và xác định các giá trị bất thường (outliers).

Kết quả thực thi:



Nhận xét tổng quan :

- Biểu đồ histogram trên cho thấy phân phối thu nhập của một khu vực cụ thể. Dựa trên hình dạng của biểu đồ, có thể rút ra một số nhận xét sau:
- **Phân phối lệch phải (right-skewed):** Phần lớn người dân trong khu vực có thu nhập trung bình hoặc dưới trung bình. Một số ít người có thu nhập rất cao, tạo nên "cái đuôi" dài về phía bên phải của biểu đồ. Điều này cho thấy sự chênh lệch thu nhập khá lớn giữa các nhóm người dân.
- **Điểm đỉnh:** Thu nhập trung bình của khu vực rơi vào khoảng 50.000 - 60.000 đơn vị tiền tệ. Đây là mức thu nhập phổ biến nhất của người dân trong khu vực.
- **Độ phân tán:** Dữ liệu thu nhập khá phân tán, thể hiện qua các cột biểu đồ có độ cao khác nhau. Điều này cho thấy sự đa dạng trong mức sống của người dân.



Nhận xét tổng quan :

- **Khoảng biến thiên:** Khoảng cách giữa giá trị nhỏ nhất và lớn nhất khá lớn, cho thấy sự phân tán của dữ liệu thu nhập khá rộng. Có một số người có thu nhập rất cao (ngoại lệ ở phía trên) và một số người có thu nhập rất thấp (ngoại lệ ở phía dưới).
- **Giá trị trung vị:** Đường ngang giữa hộp biểu thị giá trị trung vị (median), cho thấy khoảng 50% dân số có thu nhập dưới mức này và 50% còn lại có thu nhập trên mức này. Trong trường hợp này, giá trị trung vị nằm ở khoảng 60.000.
- **Các tứ phân vị:** Hộp biểu thị khoảng giữa 50% dữ liệu. Các đường thẳng ở hai đầu hộp là các tứ phân vị thứ nhất (Q1) và thứ ba (Q3). Khoảng cách giữa Q1 và Q3 cho biết khoảng biến thiên của 50% dữ liệu ở giữa.

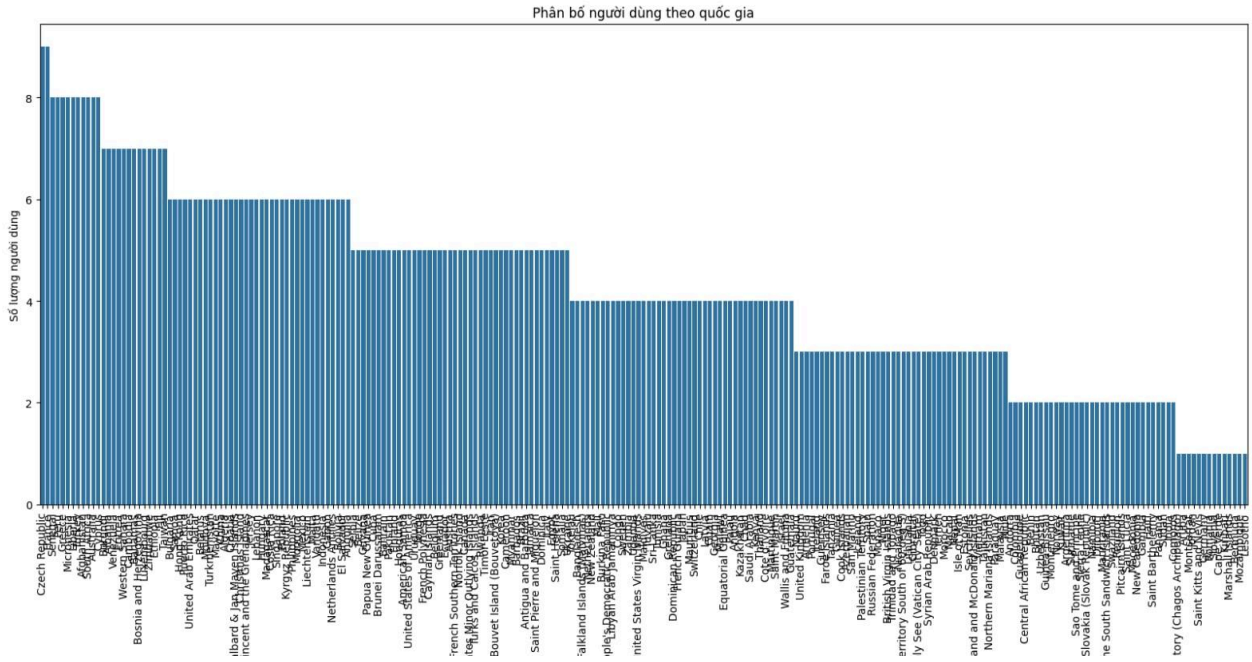
- **Các ngoại lệ:** Các điểm nằm ngoài các "râu" của biểu đồ được coi là các ngoại lệ. Trong trường hợp này, có một số ngoại lệ ở cả hai phía, cho thấy có một số người có thu nhập rất cao hoặc rất thấp so với phần còn lại của dân số.

Kết luận

- **Bất bình đẳng thu nhập:** Khu vực này có mức độ bất bình đẳng thu nhập khá cao. Một nhóm nhỏ người giàu có thu nhập cao gấp nhiều lần so với phần lớn dân số.
- **Thu nhập trung bình:** Mức thu nhập trung bình của khu vực nằm trong khoảng 50.000 - 60.000 đơn vị tiền tệ, một số ít người dùng có thu nhập rất cao hoặc rất thấp. Tuy nhiên, do sự phân tán lớn của dữ liệu, con số này có thể không phản ánh chính xác mức sống của phần lớn người dân.

f. Phân tích phân bố quốc gia (Country). Đánh giá sự phân bố người dùng theo quốc gia.

- Vẽ bar plot để quan sát số lượng người dùng theo từng quốc gia



+ Quốc gia có người dùng nhiều nhất là: Cộng hòa Séc (Czech Republic) và Pháp (France)
+ Quốc gia có người sử dụng ít nhất là: Lesotho, Mozambique, Bermuda,...

2. ***Phân tích đa biến (Bivariate Analysis):***

a. ***Quan sát mối quan hệ giữa Clicked on Ad và các biến sau:***

Gợi ý:

- **Độ tuổi**: Vẽ **boxplot** hoặc **bar plot**.
- **Thu nhập**: Dùng **scatter plot** hoặc **line plot**.
- **Giới tính**: Dùng **count plot** để quan sát tỷ lệ click theo giới tính.
- Kiểm tra sự tương quan giữa các biến độc lập và biến mục tiêu bằng **heatmap** hoặc **correlation matrix**.
- **Nhận xét** về mối quan hệ, xu hướng.

b. ***Phân tích thời gian sử dụng website theo các đặc điểm nhân khẩu học (tuổi, thu nhập, thành phố).***

Gợi ý:

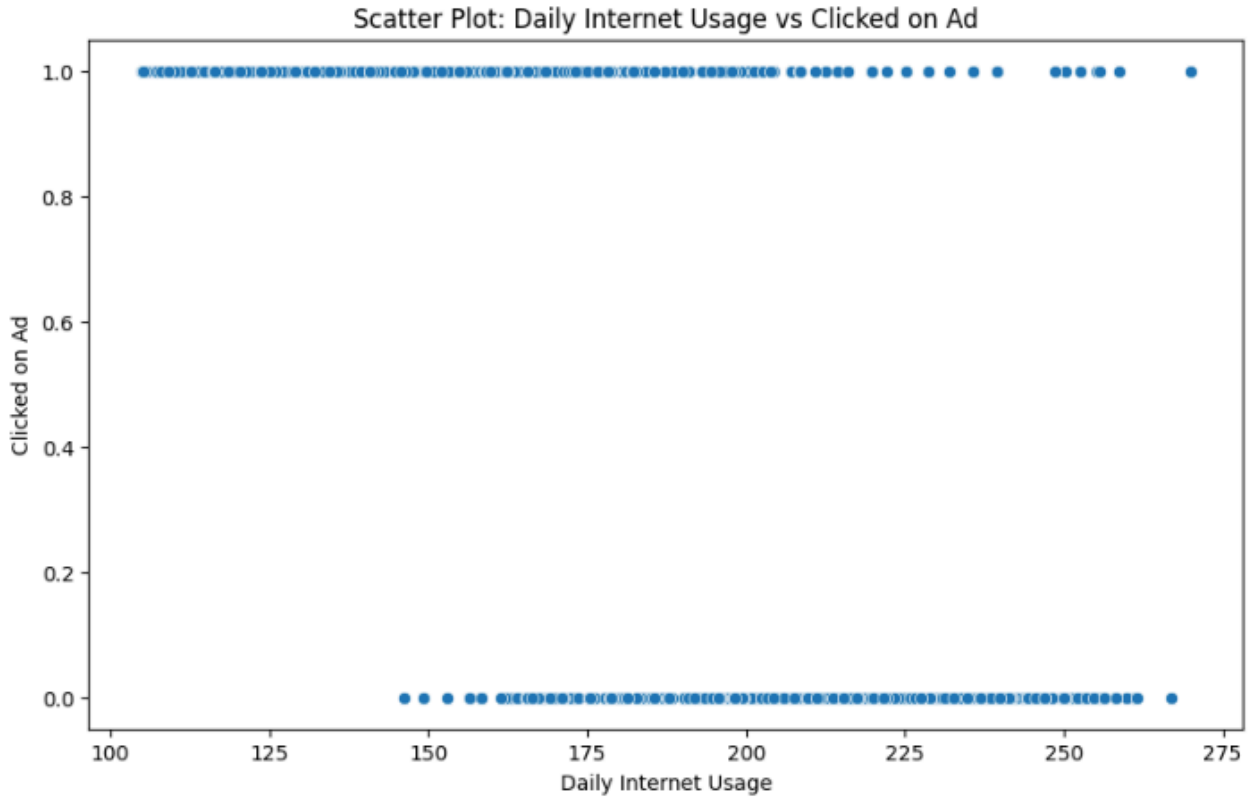
+ **Theo độ tuổi và thu nhập:**
  - Dùng scatter plot hoặc line plot.
  - Đây là phân tích hai biến (Bivariate Analysis) vì bạn đang xem mối quan hệ giữa Daily Time Spent on Site và Age hoặc Area Income.
+ **Theo thành phố:**
  - Dùng bar plot để xem phân phối thời gian sử dụng theo thành phố.

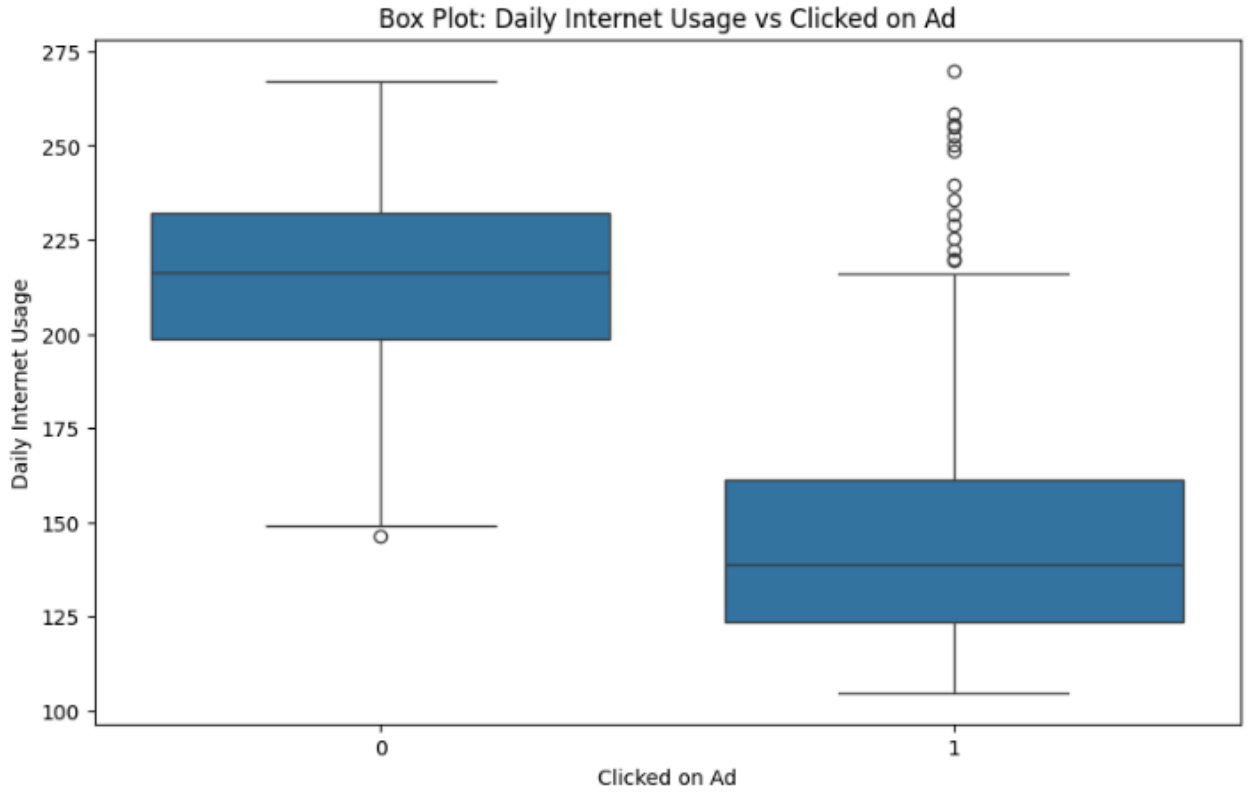
c. ***Phân tích mối quan hệ giữa thời gian sử dụng Internet và khả năng click quảng cáo (Daily Internet Usage vs Clicked on Ad)***

Gợi ý:

- Kiểm tra xem thời gian sử dụng Internet hàng ngày có ảnh hưởng đến việc click quảng cáo không.
- Sử dụng **scatter plot** hoặc **box plot** để so sánh thời gian sử dụng Internet giữa người click và không click quảng cáo.
- Nhận xét về sự khác biệt và tìm xu hướng.

Kết quả phân tích:


Scatter Plot: Daily Internet Usage vs Clicked on Ad
Clicked on Ad
Daily Internet Usage
1.0
0.8
0.6
0.4
0.2
0.0
100
125
150
175
200
225
250
275


Box Plot: Daily Internet Usage vs Clicked on Ad
Daily Internet Usage
Clicked on Ad
275
250
225
200
175
150
125
100
0
1

Nhận xét:

- Scatter Plot:
  - Scatter plot cho thấy một sự phân tán rõ ràng giữa thời gian sử dụng Internet hàng ngày và khả năng click quảng cáo.
  - Không có một xu hướng rõ ràng nào cho thấy rằng thời gian sử dụng Internet hàng ngày ảnh hưởng trực tiếp đến việc click quảng cáo. Các điểm dữ liệu phân bố khá đều giữa hai nhóm (click và không click quảng cáo).
- Box Plot:
  - Box plot cho thấy sự phân bố của thời gian sử dụng Internet giữa hai nhóm người dùng: những người click quảng cáo và những người không click quảng cáo.
  - Nhóm người dùng không click quảng cáo có thời gian sử dụng Internet hàng ngày trung bình cao hơn một chút so với nhóm người dùng click quảng cáo.
  - Tuy nhiên, sự khác biệt này không quá lớn và có thể không đủ để kết luận rằng thời gian sử dụng Internet hàng ngày ảnh hưởng đáng kể đến khả năng click quảng cáo.

Kết luận:

- Dựa trên các biểu đồ scatter plot và box plot, có thể thấy rằng thời gian sử dụng Internet hàng ngày không có ảnh hưởng rõ ràng và đáng kể đến việc click quảng cáo. Các yếu tố khác có thể đóng vai trò quan trọng hơn trong việc quyết định người dùng có click vào quảng cáo hay không.

IV. Xây Dựng Mô Hình Dự Đoán:

1. K-Nearest_Neighbors(KNN):

```
Training Accuracy : 0.73
Validation Accuracy : 0.6833333333333333

AUC: 0.6860434086461483
Time Elapsed: 0.003826141357421875 seconds
              precision    recall  f1-score   support

           0       0.64      0.79      0.71       146
           1       0.74      0.58      0.65       154

    accuracy                           0.68       300
   macro avg       0.69      0.69      0.68       300
weighted avg       0.69      0.68      0.68       300
```

- **Training Accuracy**: 0.73
- **Validation Accuracy**: 0.68
- **AUC**: 0.686
- **Precision** (Class 0): 0.64, **Recall** (Class 0): 0.79
- **Precision** (Class 1): 0.74, **Recall** (Class 1): 0.58
- **F1-Score**:
  - Class 0: 0.71
  - Class 1: 0.65
- **Time Elapsed**: 0.0038 seconds

2. Linear Regressor:

```
Logistic Regression Accuracy (all features): 0.8933333333333333
Classification Report for Logistic Regression (all features):
              precision    recall  f1-score   support

           0       0.87      0.92      0.89       146
           1       0.92      0.87      0.89       154

    accuracy                           0.89       300
   macro avg       0.89      0.89      0.89       300
weighted avg       0.89      0.89      0.89       300
```

- **Accuracy**: 0.89

- **Precision** (Class 0): 0.87, **Recall** (Class 0): 0.92
- **Precision** (Class 1): 0.92, **Recall** (Class 1): 0.87
- **F1-Score**: 0.89 cho cả hai lớp
- **Classification Report**:
    - **Precision** và **Recall** đều cao cho cả hai lớp, cho thấy mô hình phân loại rất chính xác và phát hiện tốt cả hai lớp.

Đánh giá: **Logistic Regression** cho **accuracy cao nhất** (0.89) trong ba mô hình, với **precision** và **recall** khá cân bằng cho cả hai lớp. Mô hình này thực hiện tốt trong việc phân loại **click ads** và **non-click ads**.
F1-Score ở mức **0.89** cho cả hai lớp là một chỉ số tốt, cho thấy mô hình có sự cân bằng giữa việc phát hiện các đối tượng và không đưa ra nhiều dự đoán sai.

3. Random Forest:

```
Random Forest Classifier Accuracy: 0.89

Random Forest Classifier Classification Report:
              precision    recall  f1-score   support

           0       0.86      0.92      0.89       146
           1       0.92      0.86      0.89       154

    accuracy                           0.89       300
   macro avg       0.89      0.89      0.89       300
weighted avg       0.89      0.89      0.89       300
```

- **Accuracy**: 0.89
- **Precision** (Class 0): 0.86, **Recall** (Class 0): 0.92
- **Precision** (Class 1): 0.92, **Recall** (Class 1): 0.86
- **F1-Score**: 0.89 cho cả hai lớp
- **Classification Report**:
    - **Random Forest** có độ chính xác và khả năng phát hiện tốt cho cả hai lớp. Mô hình này cũng cho **precision** và **recall** rất cân bằng cho cả **Class 0** và **Class 1**.

Đánh giá: **Random Forest Classifier** có kết quả tương đương **Logistic Regression** về **accuracy** và các độ đo khác. Cả hai mô hình đều có **accuracy** 0.89 và **F1-Score** 0.89 cho cả hai lớp.

Tuy nhiên, **Random Forest** có thể có lợi thế về khả năng xử lý các mối quan hệ phi tuyến tính phức tạp hơn, nhưng với dữ liệu hiện tại, nó không thể hiện sự khác biệt rõ ràng so với **Logistic Regression**.

V. Đánh Giá Mô Hình

1. K-Nearest_Neighbors(KNN):

KNN có **accuracy thấp** trên tập kiểm tra (0.68), cho thấy mô hình có thể gặp khó khăn trong việc phân loại chính xác các đối tượng, đặc biệt là đối với **Class 1** (click ads). Mặc dù **precision** và **recall** của **Class 1** khá tốt, nhưng **recall** thấp của **Class 1** chỉ ra rằng mô hình không phát hiện đầy đủ các trường hợp **click ads**.

Mặc dù vậy, mô hình **KNN** vẫn cho thấy khả năng phân loại tốt đối với **Class 0** (không click ads), với **recall** cao.

2. Linear Regressor:

Logistic Regression cho **accuracy cao nhất** (0.89) trong ba mô hình, với **precision** và **recall** khá cân bằng cho cả hai lớp. Mô hình này thực hiện tốt trong việc phân loại **click ads** và **non-click ads**.

F1-Score ở mức **0.89** cho cả hai lớp là một chỉ số tốt, cho thấy mô hình có sự cân bằng giữa việc phát hiện các đối tượng và không đưa ra nhiều dự đoán sai.

3. ***Random Forest:***

   Random Forest Classifier có kết quả tương đương **Logistic Regression** về **accuracy** và các độ đo khác. Cả hai mô hình đều có **accuracy** 0.89 và **F1-Score** 0.89 cho cả hai lớp.

   Tuy nhiên, **Random Forest** có thể có lợi thế về khả năng xử lý các mối quan hệ phi tuyến tính phức tạp hơn, nhưng với dữ liệu hiện tại, nó không thể hiện sự khác biệt rõ ràng so với **Logistic Regression**.

4. ***Kết luận:***
   - **Logistic Regression** và **Random Forest** đều cho kết quả **accuracy cao nhất** (0.89) và **F1-Score** 0.89, với **precision** và **recall** cân bằng tốt cho cả hai lớp.
   - **KNN** có **accuracy thấp hơn** (0.68) và **F1-Score** thấp hơn, đặc biệt với **Class 1**, cho thấy nó không phải là sự lựa chọn tốt nhất trong trường hợp này.
   - Nếu muốn một mô hình đơn giản và dễ triển khai, **Logistic Regression** có thể là sự lựa chọn tốt vì nó đạt được kết quả cao mà không yêu cầu quá nhiều tài nguyên tính toán.
   - **Random Forest** có thể phù hợp nếu cần một mô hình linh hoạt và có thể phát hiện các mối quan hệ phi tuyến tính phức tạp hơn, mặc dù trong trường hợp này, kết quả của nó không khác biệt đáng kể so với **Logistic Regression**.

Tóm lại, **Logistic Regression** và **Random Forest** đều là lựa chọn tốt, và quyết định nên dựa trên các yếu tố như tính toán tài nguyên, tính khả thi của việc triển khai, và yêu cầu về độ phức tạp của mô hình.

VI. Kết luận:

Đồ án thực hiện một nghiên cứu rất có giá trị trong việc phân tích hành vi người dùng và xây dựng mô hình dự đoán khả năng click quảng cáo. Việc đánh giá mô hình qua các độ đo như **precision**, **recall**, và **F1-score** đã giúp chọn ra mô hình tối ưu và cung cấp cái nhìn sâu sắc về những yếu tố ảnh hưởng đến khả năng click quảng cáo. Mặc dù mô hình **Logistic Regression** và **Random Forest** có kết quả gần tương đương, nhưng **Logistic Regression** có thể là sự lựa chọn tốt trong trường hợp này do tính đơn giản và hiệu quả cao.

VII. Đánh giá thực hiện nghiên cứu của nhóm:

1. *Ưu điểm và nhược điểm của nhóm:*

- **Làm việc nhóm**: Các thành viên trong nhóm đã phối hợp tốt trong việc chia sẻ công việc và thực hiện các nhiệm vụ liên quan đến EDA, xây dựng mô hình và đánh giá mô hình.
- **Chia sẻ kiến thức**: Các thành viên đã áp dụng các kỹ thuật thống kê, phân tích và học máy đúng cách, giúp xây dựng một đồ án khoa học và có tính ứng dụng cao.
- Mọi bước trong quy trình được làm rõ ràng với các báo cáo, kết quả và mã nguồn được chia sẻ minh bạch trong file .ipynb và link Colab, dễ dàng kiểm tra và tái tạo.

2. *Nhược điểm của nhóm:*

- **Thiếu sự đa dạng trong mô hình**: Mặc dù đã thử một số mô hình (KNN, Logistic Regression, Random Forest), nhưng có thể thêm vào các mô hình khác để đánh giá và cải thiện kết quả, chẳng hạn như SVM, Gradient Boosting hay XGBoost.
- **Thiếu một số kỹ thuật tiền xử lý nâng cao**: Mặc dù xử lý missing values và outliers đã được thực hiện, nhưng có thể áp dụng thêm các kỹ thuật như SMOTE cho vấn đề mất cân bằng dữ liệu hoặc tối ưu hóa mô hình qua các phương pháp như Grid Search hay Random Search.

3. *Mức độ phổ biến trong các dự án:*

Dự án này là một ví dụ điển hình của bài toán phân tích và dự đoán hành vi người dùng trong các hệ thống quảng cáo trực tuyến, một ứng dụng phổ biến trong các dự án dữ liệu lớn và tiếp thị số. Các phương pháp và kỹ thuật trong đồ án, như **Exploratory Data Analysis (EDA)**, **tiền xử lý dữ liệu**, **xây dựng mô hình học máy**, và **đánh giá mô hình** là rất phổ biến và có thể áp dụng vào nhiều dự án trong các lĩnh vực khác trong thực tế.

PHÂN CÔNG CÔNG VIỆC

STT	Công việc	Người thực hiện	Ngày hoàn thành
GIAI ĐOẠN 1 : Data Preparation + Data Preprocessing			
1	Tạo template Báo cáo, phân công	Lô Thủy Tiên	00:00 AM 26/10/2024
2	Tìm hiểu lý thuyết: ***GIỚI THIỆU TỔNG QUÁT : ĐỒ ÁN THỰC HÀNH #1***	Lô Thủy Tiên	00:00 AM 28/10/2024
3	THỰC HIỆN ĐỒ ÁN : ***Chuẩn bị dữ liệu (Data Preparation)***	Lô Thủy Tiên	00:00 AM 28/10/2024
4	THỰC HIỆN ĐỒ ÁN : ***Khám phá và tiền xử lý dữ liệu (Data Preprocessing)*** + Handling Outliers | Non Relevant Data + Incorrect Types + Standardizing Data	Nguyễn Vũ Tường An	00:00 AM 03/11/2024
5	THỰC HIỆN ĐỒ ÁN : ***Khám phá và tiền xử lý dữ liệu (Data Preprocessing):*** + Handling Missing Value & Empty Data + Handling Duplications | Errors	Lô Thủy Tiên	00:00 AM 03/11/2024
6	THỰC HIỆN ĐỒ ÁN : ***Khám phá và tiền xử lý dữ liệu (Data Preprocessing):*** + Incorrect | Invalid Values + Normalizing Data + Validating Data	Lê Mạnh Hoàng	00:00 AM 03/11/2024

GIAI ĐOẠN 2 : EDA			
6	***Phân tích đơn biến (Univariate Analysis), Bivariate Analysis:*** Kiểm tra sự cân bằng giữa nhãn: click/không click quảng cáo + Sử dụng value_counts() để đếm số lượng của từng nhãn (Clicked on Ad). + Trực quan hóa với biểu đồ cột (countplot) để dễ quan sát sự mất cân bằng. + Cap màn hình kết quả + nhận xét trong báo cáo.	Lê Mạnh Hoàng	00:00 AM 29/11/2024
7	***Phân tích đơn biến (Univariate Analysis), Bivariate Analysis:*** Phân tích các biến số (age, income, gender, daily internet usage, etc.). + Tính toán thống kê cơ bản: trung bình, độ lệch chuẩn, phân phối. + Vẽ biểu đồ: ● **Biểu đồ histogram** cho các biến số liên tục như độ tuổi (age), thu nhập (income), thời gian sử dụng (daily_time_spent). ● **Biểu đồ bar plot** cho các biến phân loại như giới tính (Male).	Nguyễn Vũ Tường An	00:00 AM 29/11/2024

	+ Cap màn hình kết quả + nhận xét trong báo cáo.		
8	***Phân tích đơn biến (Univariate Analysis), Bivariate Analysis:*** Phân tích thời điểm sử dụng website trong ngày + Phân chia thành các khung giờ (Sáng, Trưa, Chiều, Đêm). + Vẽ bar plot hoặc count plot để quan sát phân bố số lượng người dùng. + Cap màn hình kết quả + nhận xét trong báo cáo.	Lô Thủy Tiên	00:00 AM 29/11/2024
9	***Phân tích đơn biến (Univariate Analysis), Bivariate Analysis:*** Phân tích chủ đề quảng cáo. Xác định các chủ đề quảng cáo phổ biến. + Tạo nhóm chủ đề từ Ad Topic Line. + Dùng count plot hoặc word cloud để xem các chủ đề quảng cáo. + Cap màn hình kết quả + nhận xét trong báo cáo.	Nguyễn Vũ Tường An	00:00 AM 29/11/2024
10	***Phân tích đơn biến (Univariate Analysis), Bivariate Analysis:*** Phân tích thu nhập khu vực (Area Income). Đánh giá phân phối thu nhập trung bình của người dùng. + Vẽ **histogram** hoặc **box plot** cho cột **Area Income**.	Lô Thủy Tiên	00:00 AM 29/11/2024

	+ Quan sát sự phân bố thu nhập và xác định các giá trị bất thường (outliers). + Cap màn hình kết quả + nhận xét trong báo cáo.		
11	***Phân tích đơn biến (Univariate Analysis), Bivariate Analysis:*** Phân tích phân bố quốc gia (Country). Đánh giá sự phân bố người dùng theo quốc gia. + Vẽ bar plot để quan sát số lượng người dùng theo từng quốc gia. + Xác định quốc gia có số lượng người dùng nhiều nhất và ít nhất. + Cap màn hình kết quả + nhận xét trong báo cáo.	Lê Mạnh Hoàng	00:00 AM 29/11/2024
12	***Phân tích hai biến (Bivariate Analysis):*** + Quan sát mối quan hệ giữa Clicked on Ad và các biến sau: • **Độ tuổi**: Vẽ **boxplot** hoặc **bar plot**. • **Thu nhập**: Dùng **scatter plot** hoặc **line plot**. • **Giới tính**: Dùng **count plot** để quan sát tỷ lệ click theo giới tính.	Lô Thủy Tiên	00:00 AM 29/11/2024

	+ Kiểm tra sự tương quan giữa các biến độc lập và biến mục tiêu bằng **heatmap** hoặc **correlation matrix**. + **Nhận xét** về mối quan hệ, xu hướng.		
13	***Phân tích hai biến (Bivariate Analysis):*** Phân tích thời gian sử dụng website theo các đặc điểm nhân khẩu học (tuổi, thu nhập, thành phố). + **Theo độ tuổi và thu nhập:** ● Dùng scatter plot hoặc line plot. ● Đây là phân tích hai biến (Bivariate Analysis) vì bạn đang xem mối quan hệ giữa Daily Time Spent on Site và Age hoặc Area Income. + **Theo thành phố:** ● Dùng bar plot để xem phân phối thời gian sử dụng theo thành phố.	Lê Mạnh Hoàng	00:00 AM 29/11/2024
14	***Phân tích hai biến (Bivariate Analysis):*** Phân tích mối quan hệ giữa thời gian sử dụng Internet và khả năng click quảng cáo (Daily Internet Usage vs Clicked on Ad)	Nguyễn Vũ Tường An	00:00 AM 29/11/2024

	+ Kiểm tra xem thời gian sử dụng Internet hàng ngày có ảnh hưởng đến việc click quảng cáo không. + Sử dụng **scatter plot** hoặc **box plot** để so sánh thời gian sử dụng Internet giữa người click và không click quảng cáo. + Nhận xét về sự khác biệt và tìm xu hướng.		
GIAI ĐOẠN 3 : Xây dựng mô hình dự đoán + Đánh giá mô hình			
15	**K-Nearest Neighbors (kNN)**	Lô Thủy Tiên	00:00 AM 12/12/2024
16	**Random Forest**	Nguyễn Vũ Tường An	00:00 AM 12/12/2024
17	**Logistic Regression**	Lê Mạnh Hoàng	00:00 AM 12/12/2024

ĐÁNH GIÁ THÀNH VIÊN

Nhóm : 03

MSSV	Họ tên	Hoàn thành	Nhận xét
18127008	Lê Mạnh Hoàng	100%	
21127211	Nguyễn Vũ Tường An	100%	
21127699	Lô Thủy Tiên	100%	

TÀI LIỆU THAM KHẢO

Công cụ và phần mềm hỗ trợ:

STT	Chức năng	Công cụ
[1]	Code	Microsoft Visual Studio 2022, VS code
[2]	Báo cáo	Google Docs
[3]	Quản lý, trao đổi	Facebook, Messenger
[4]	Họp định kỳ	Google Meet
[5]	AI support	ChatGPT, Copilot, Perplexity, Gemini

Tài liệu tham khảo:

[1] https://www.mcivietnam.com/blog-detail/kham-pha-05-du-an-phan-tich-du-lieu-thu-vi-cho-nguoi-moi-bat-au-GG7DB7/

[2] https://www.youtube.com/watch?v=DJofs2JyIVM

[3] https://www.youtube.com/watch?v=VH2JgqlN2so

[4] https://github.com/lhminhtuan2000/PTDLUD_DATH_2

[5] Datapot. (n.d.). Các loại biểu đồ trong Power BI – Phần 1. Truy cập từ: https://datapot.vn/cac-loai-bieu-do-trong-power-bi-phan-1/ 1

[6] vngson. (n.d.). *PTDLUD*. Truy cập từ: https://github.com/vngson/PTDLUD/blob/main/main.ipynb 2

[7] HomelessSandwich. (n.d.). *Analysis of Advertising Data*. Truy cập từ: https://www.kaggle.com/code/homelesssandwich/analysis-of-advertising-data#Building-Models-on-Data

[8] Ahmed Abdellah Ismail. (n.d.). *Binary Classification*. Truy cập từ: https://www.kaggle.com/code/ahmedabdellahismail/binary-classification#Exploratory-Data-Analysis

[9] Gabriel Santello. (n.d.). *Advertisement Click on Ad*. Truy cập từ: https://www.kaggle.com/datasets/gabrielsantello/advertisement-click-on-ad/code?datasetId=2311024&searchQuery=k-nn

[10] YouTube. (2020). *Video Title: Machine learning | Mô hình phân lớp kNN (k-Nearest Neighbors)*. Truy cập từ: https://www.youtube.com/watch?v=ek0uoghdDaY