BÁO CÁO TIỀN XỬ LÝ DỮ LIỆU

Họ và tên: Đinh Bùi Thu Linh

MSV: SIC0082

Lớp: TL - HN AI2

1. Thông tin về bộ dữ liệu

Bộ dữ liệu ad_ctr.csv bao gồm 10000 dòng với các trường thông tin như sau:

Daily Time Spent on Site	Thời gian người dùng truy cập trang web tính bằng phút
Age	Tuổi của người dùng
Area income	Thu nhập trung bình của khu vực địa lý của người tiêu dùng
Daily Internet Usage	Trung bình số phút trong ngày người dùng sử dụng internet
Ad Topic Line	Tiêu đề bài quảng cáo
City	Thành phố của người dùng
Male	Giới tính của người dùng (Nam = có / Nữ = không)
Country	Quốc gia của người dùng
Timestamp	Thời gian người dùng click vào quảng cáo hoặc đóng quảng cáo
Clicked on Ad	Người dùng có Click vào quảng cáo hay không (có = 1 / không = 0)

2. Tiền xử lý dữ liệu

2.1. Kiểm tra giá trị rỗng, trùng lặp

```
df.duplicated().sum()
```

```
215
```

Dữ liệu có 215 giá trị bị trùng lặp, tiến hành loại bỏ các giá trị này

```
df.drop_duplicates(inplace = True)
df.duplicated().sum()
```

```
0
```

Tiếp theo kiểm tra xem dữ liệu có giá trị rỗng không

```
df.isnull().sum()
```

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

Như hình ta thấy dữ liệu không có giá trị rỗng, nên bỏ qua các bước xử lý giá trị rỗng đến các bước xử lý tiếp theo.

2.2. Xử lý các giá trị numerical và categorical

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 10 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   Daily Time Spent on Site  10000 non-null  float64
 1   Age                       10000 non-null  float64
 2   Area Income               10000 non-null  float64
 3   Daily Internet Usage      10000 non-null  float64
 4   Ad Topic Line             10000 non-null  object 
 5   City                      10000 non-null  object 
 6   Gender                    10000 non-null  object 
 7   Country                   10000 non-null  object 
 8   Timestamp                 10000 non-null  object 
 9   Clicked on Ad             10000 non-null  int64  
dtypes: float64(4), int64(1), object(5)
```

Như vậy dữ liệu có các cột numeric là: 'Daily Time Spent on Site', 'Age', 'Area Income', 'Daily Internet Usage' và các cột category là: 'Ad Topic Line', 'City', 'Gender', 'Country', 'Clicked on Ad'.

Đối với các dữ liệu dạng numeric ta có bảng mô tả như sau:

```
df[numeric_cols].describe()
```

	Daily Time Spent on Site	Age	Area Income	Daily Internet Usage
count	9785.000000	9785.000000	9785.000000	9785.000000
mean	61.601379	35.839550	53948.143348	177.886144
std	15.698216	8.538524	13360.051625	40.861875
min	32.600000	19.000000	13996.500000	105.220000
25%	48.030000	29.000000	44174.250000	140.150000
50%	59.590000	35.000000	56180.930000	178.920000
75%	76.270000	41.000000	62669.590000	212.870000
max	90.970000	60.000000	79332.330000	269.960000

Ta thấy giá trị mean và median(50%) của mỗi cột đều xấp xỉ bằng nhau nên chúng ta có thể bỏ qua quá trình xử lý cân bằng dữ liệu.

Với các cột dạng categorical ta có các giá trị riêng biệt trong mỗi cột như sau:

```
for i in ((Categorical_cols)) :
    print(i)
    print(len(df[i].unique()))
    print()
```

```
Ad Topic Line
559

City
521

Gender
2

Country
207

Clicked on Ad
2
```

```
df[Categorical_cols].describe(include = ['O'])
```

	Ad Topic Line	City	Gender	Country
count	9785	9785	9785	9785
unique	559	521	2	207
top	Cloned explicit middleware	Hubbardmouth	Female	Australia
freq	323	330	5268	346

Ta thấy ở đây có quá nhiều thành phố và không có nhiều người thuộc cùng một thành phố. Nên có khả năng cao cột City sẽ không được sử dụng trong bài toán dự đoán. Cột Country có thể để lại để xem xét thêm

2.3. Xử lý giá trị thời gian

Với cột Timestamp ta hoàn toàn có thể xử lý chúng bằng cách biến đổi thành các giá trị Hour, Day of Week, Date, Month. Mục đích của việc xử lý này là có thể giúp nhìn nhận và phân tích lưu lượng truy cập của người dùng chi tiết

Dữ liệu sau khi xử lý có dạng như sau:

```
df.head()
```

	Daily Time Spent on Site	Age	Area Income	Daily Internet Usage	Ad Topic Line	City	Gender	Country	Clicked on Ad	Month	Day	Hour	Weekday	Date
0	62.26	32.0	69481.85	172.83	Decentralized real-time circuit	Lisafort	Male	Svalbard & Jan Mayen Islands	0	6	9	21	3	2016-06-09
1	41.73	31.0	61840.26	207.17	Optional full-range projection	West Angelabury	Male	Singapore	0	1	16	17	5	2016-01-16
2	44.40	30.0	57877.15	172.83	Total 5thgeneration standardization	Reyesfurt	Female	Guadeloupe	0	6	29	10	2	2016-06-29
3	59.88	28.0	56180.93	207.17	Balanced empowering success	New Michael	Female	Zambia	0	6	21	14	1	2016-06-21
4	49.21	30.0	54324.73	201.58	Total 5thgeneration standardization	West Richard	Female	Qatar	1	7	21	10	3	2016-07-21

2.4. Xử lý giá trị ngoại lệ

Sử dụng IQR để xử lý các giá trị ngoại lệ. Đối với bất kỳ biến định lượng nào, các điểm lớn hơn 1,5 IQR ở trên hoặc dưới các phần tư trên và dưới được giả định là các ngoại lệ.

Dưới đây là bảng thống kê các giá trị của các cột "Daily Time Spent on Site,Age,Area Income, Daily Internet Usage"

count 9785.000000 mean 61.601379 std 15.698216 min 32.600000 25% 48.030000 50% 59.590000 75% 76.270000 max 90.970000 Name: **Daily Time Spent on Site**, dtype: float64 Không có ngoại lệ ở cột: Daily Time Spent on Site Ngoại lệ dưới: 5.670000000000009 Ngoại lệ trên: 118.63	count 9785.000000 mean 35.839550 std 8.538524 min 19.000000 25% 29.000000 50% 35.000000 75% 41.000000 max 60.000000 Name: **Age**, dtype: float64 Có ngoại lệ ở cột: Age Ngoại lệ dưới: 11.0 Ngoại lệ trên: 59.0
count 9785.000000 mean 53948.143348 std 13360.051625 min 13996.500000 25% 44174.250000 50% 56180.930000 75% 62669.590000 max 79332.330000 Name: **Area Income,** dtype: float64 Có ngoại lệ ở cột: Area Income Ngoại lệ dưới: 16431.240000000005 Ngoại lệ trên: 90412.59999999999	count 9785.000000 mean 177.886144 std 40.861875 min 105.220000 25% 140.150000 50% 178.920000 75% 212.870000 max 269.960000 Name: **Daily Internet Usage**, dtype: float64 Không có ngoại lệ ở cột: Daily Internet Usage Ngoại lệ dưới: 31.070000000000007 Ngoại lệ trên: 321.95

Số dòng có ngoại lệ: 82

Xét một số hàng có giá trị ngoại lệ

	Daily Time Spent on Site	Age	Area Income	Daily Internet Usage	Ad Topic Line	City	Gender	Country	Clicked on Ad	Month	Day	Hour	Weekday	Date
223	56.39	60.0	69646.35	218.61	Programmable uniform website	Carterland	Female	Belgium	1	3	16	20	2	2016-03-16
447	47.64	60.0	69646.35	186.37	Programmable uniform website	Lisamouth	Female	Australia	1	3	9	0	2	2016-03-09
932	51.87	60.0	51067.54	119.86	Polarized 5thgeneration matrix	Hansenmouth	Female	Peru	1	4	20	10	2	2016-04-20
970	59.51	60.0	40468.53	218.61	Polarized 5thgeneration matrix	Youngfort	Female	Antigua and Barbuda	1	2	27	12	5	2016-02-27
1073	59.51	60.0	58966.22	153.76	Universal empowering adapter	Williamsside	Female	Australia	1	1	5	16	1	2016-01-05

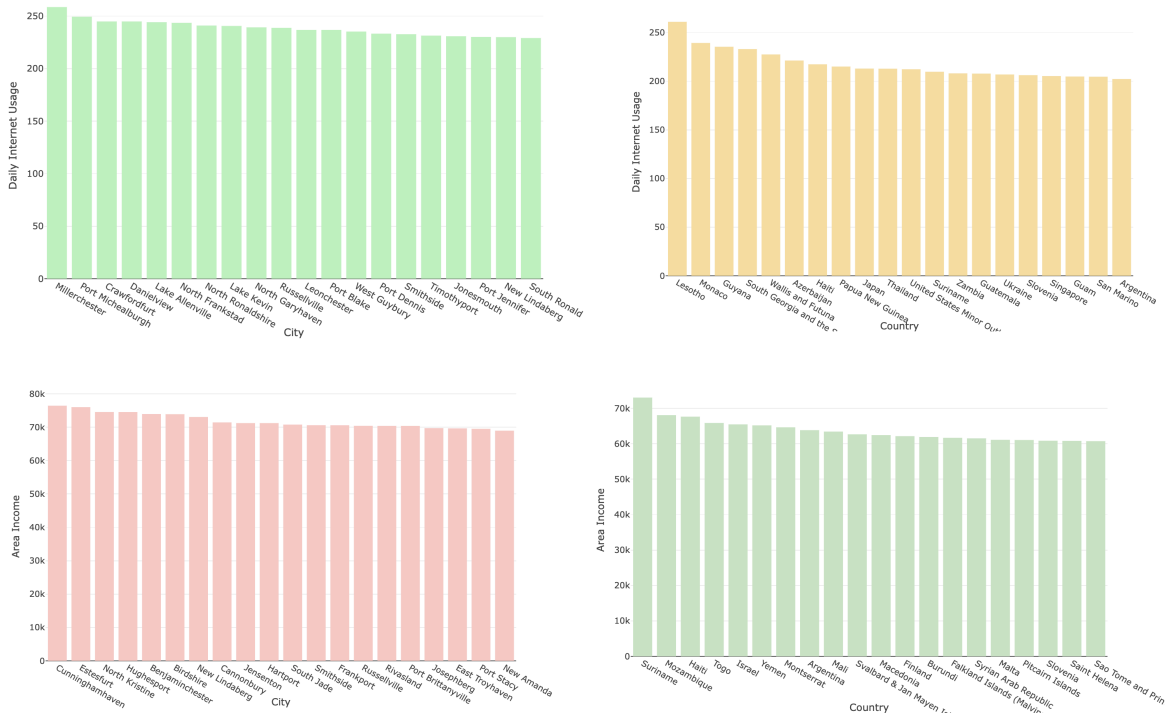
Tất cả những người này đều thuộc về biến Area Income. Vì lý do này, chúng tôi sẽ không sắp xếp những người này là ngoại lệ vì họ có thể đến từ các khu vực có thu nhập thấp.

Hơn nữa, trừ khi nguồn dữ liệu thu nhập khu vực không chính xác hoặc có sự hiểu lầm về cách biến đó được thu thập/tạo ra, nếu không thì không cần phải loại bỏ những người này.

3. Trực quan hoá dữ liệu

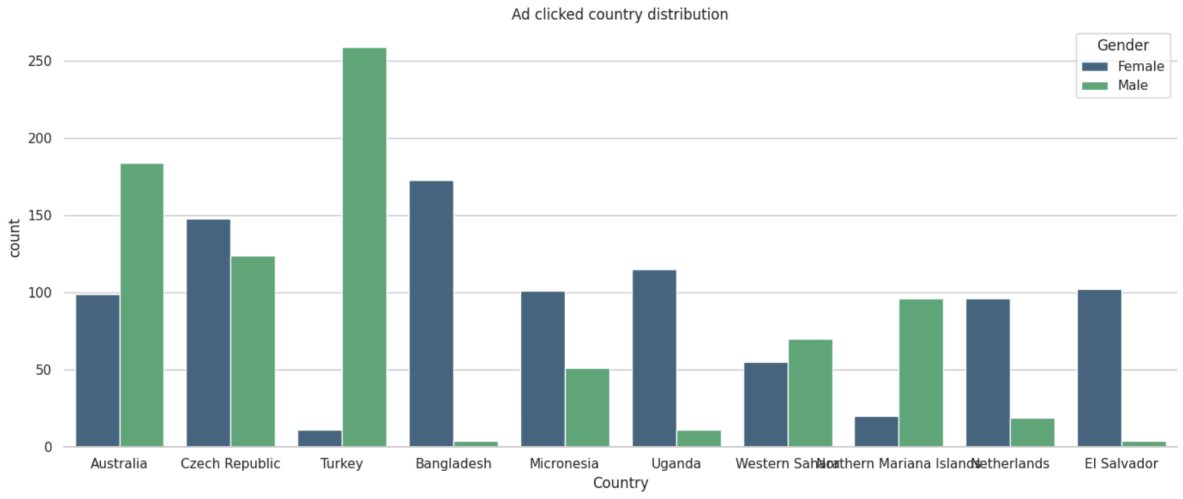
Trực quan hoá dữ liệu là một phương pháp hữu hiệu để giúp ta có thể hiểu dữ liệu của mình hơn. Để thực hiện công việc này trước hết ta đặt ra một số câu hỏi liên quan đến dữ liệu.

Đầu tiên thực hiện việc xem xét chi tiết cột City và Country có nên giữ lại trong bộ dữ liệu hay không?

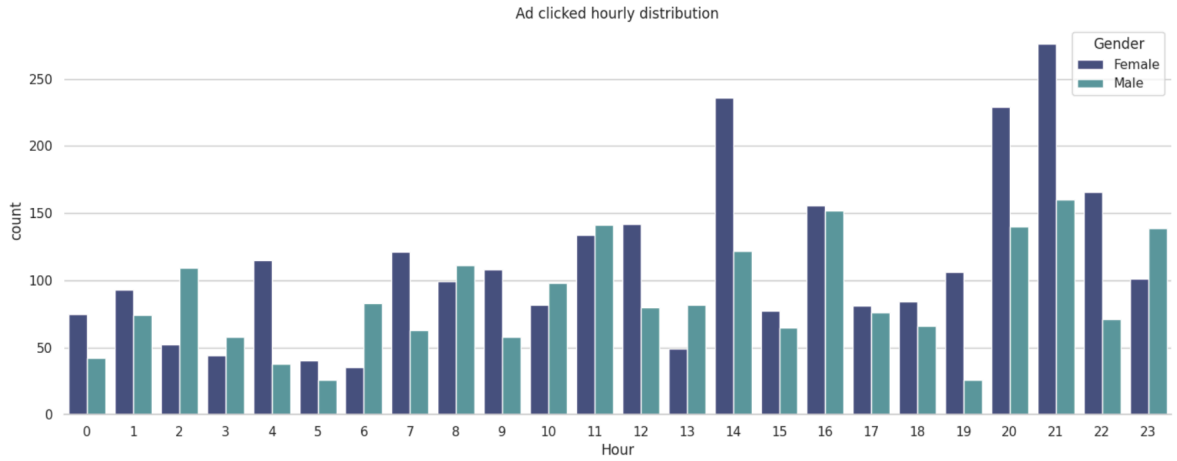


Hình trên cho thấy top 20 thành phố và quốc gia trong các thông tin Trung bình thời gian sử dụng Internet và thu nhập đầu vào. Trái với suy nghĩ ban đầu là các thành phố lớn hay các quốc gia lớn, phát triển sẽ có lượng người dùng truy cập lớn. Ở đây top 20 của mỗi thành phần không hề giống nhau về thứ tự. Chứng tỏ rằng lượng người dùng truy cập vào quảng cáo gần như không hề phụ thuộc vào cột City và cột Country. Vì vậy nên khi dự đoán có thể loại bỏ hai cột này

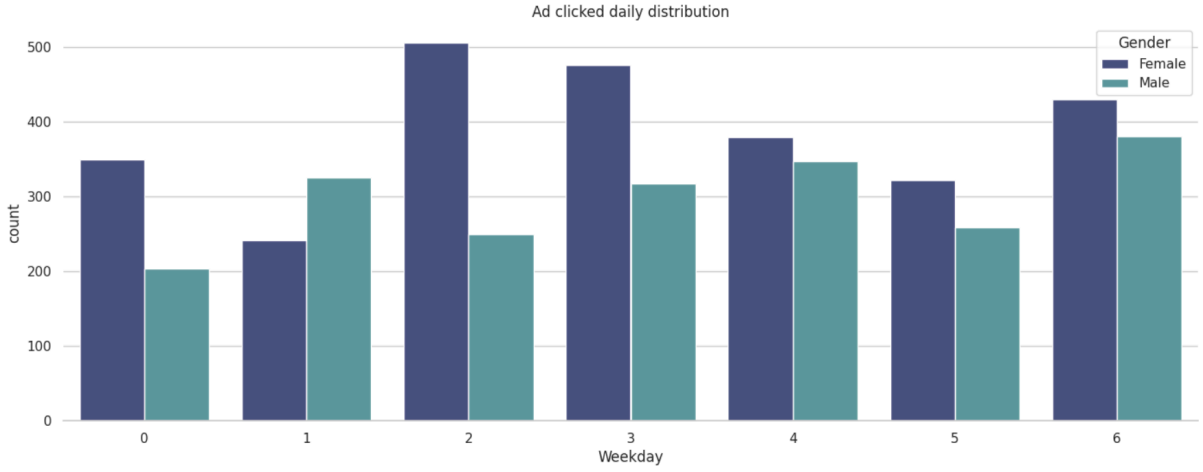
Tiếp đến ta sẽ xem phân bố lượng click quảng cáo của 10 quốc gia hàng đầu dựa vào giới tính



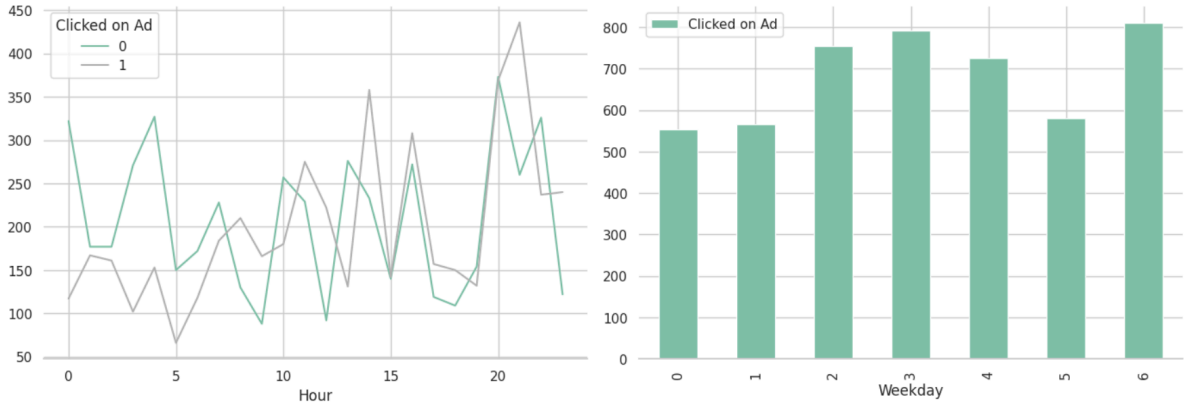
Nhìn vào có thể thấy các nước phát triển thì nữ giới đóng góp lượng click lớn hơn so với nam giới.



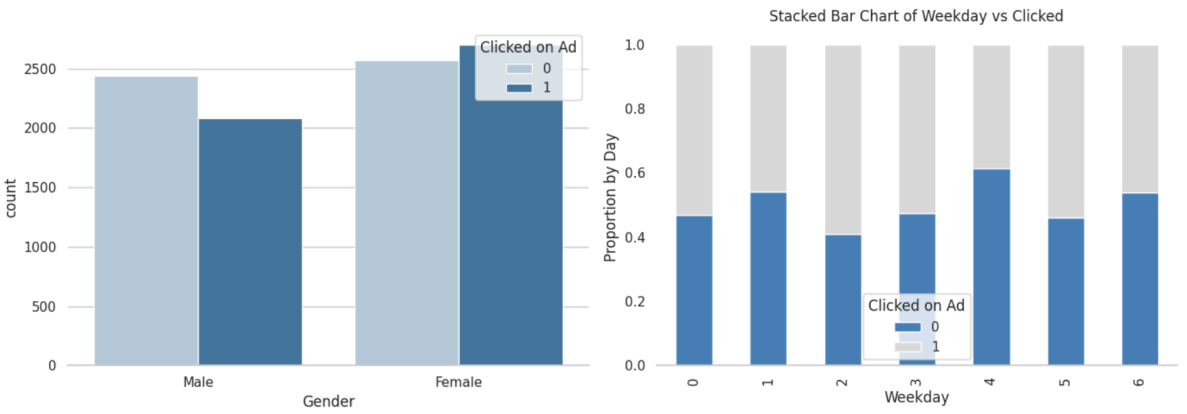
Với phân bổ theo giờ, ta có thể thấy lượng truy cập của nữ giới vẫn cao hơn hẳn nam giới, đặc biệt ở 14h, 20h,21h. Như vậy nếu muốn tăng hiệu quả của quảng cáo, nhất là các quảng cáo dành cho phái đẹp, nên ưu tiên các giờ trên để đẩy quảng cáo.



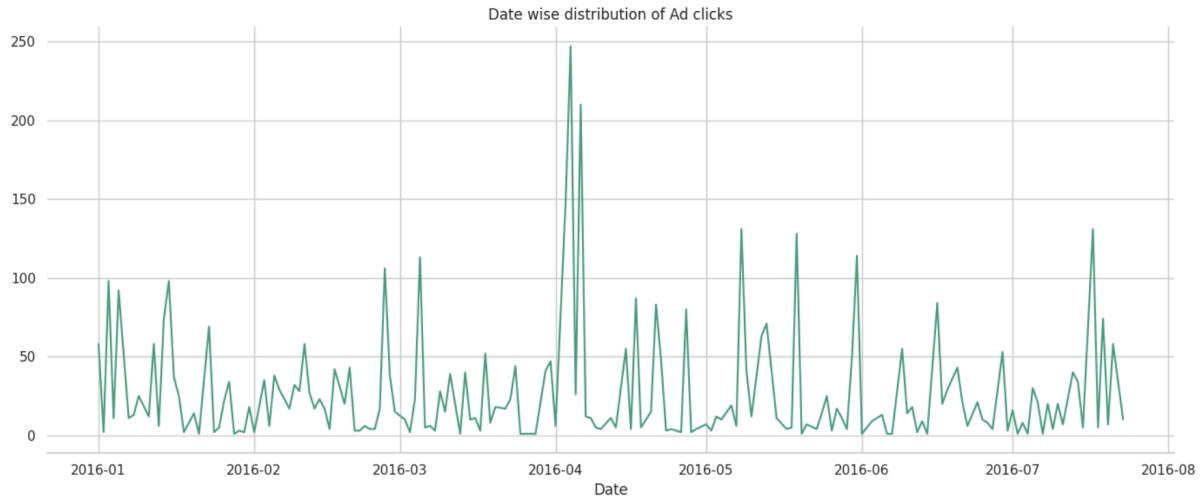
Không chỉ có thế, với biểu đồ theo ngày trong tuần, nữ giới có xu hướng xem quảng cáo nhiều vào thứ tư, thứ năm. Còn nam giới có xu hướng xem nhiều thêm vào các ngày thứ sáu, chủ nhật.



Biểu đồ đường ở đây cho biết rằng người dùng có xu hướng nhấp vào Quảng cáo vào cuối ngày hoặc có thể là vào sáng sớm. Điều này được mong đợi dựa trên đặc điểm độ tuổi mà hầu hết mọi người đang làm việc, vì vậy nó có vẻ phù hợp khi họ tìm thấy thời gian sớm hoặc muộn trong ngày. Ngoài ra, Chủ nhật dường như có hiệu quả khi nhấp vào quảng cáo từ biểu đồ thanh.



Nhìn vào biểu đồ xếp chồng ta có thể thấy thời điểm tốt nhất trong tuần có lượt click cao nhất là Thứ Năm.



Xét biểu đồ thời gian, ta thấy tháng 4 năm 2016 có lượng click cao nhất

Date

2016-04-04 247

2016-04-06 210

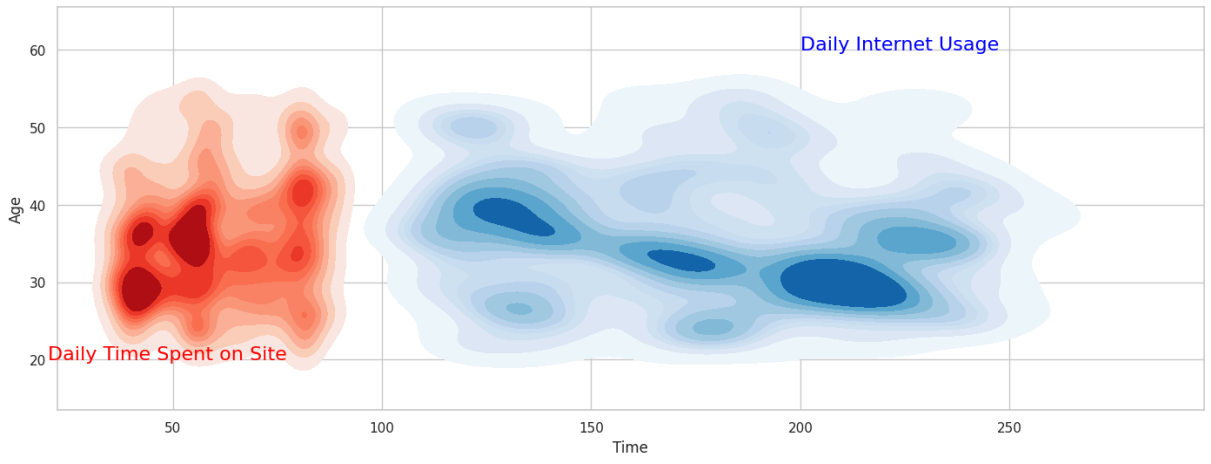
2016-04-03 148

2016-05-08 131

2016-07-17 131

Name: count, dtype: int64

Như vậy với dữ liệu quảng cáo này thì thời gian quảng cáo hiệu quả nhất là Tháng 4.



Như chúng ta có thể thấy, những người ở độ tuổi khoảng 30 dành nhiều thời gian cho Internet và trang web, nhưng họ không nhấp vào Quảng cáo thường xuyên. So với họ, dân số khoảng 40 tuổi dành ít thời gian hơn một chút nhưng lại nhấp vào Quảng cáo nhiều hơn.

KẾT LUẬN

1. Kết quả đạt được

- Thực hiện các bước tiền xử lý dữ liệu cơ bản như loại bỏ giá trị rỗng, ngoại lệ, thời gian,...
- Thực hiện các bước trực quan hoá dữ liệu để hiểu sâu hơn về dữ liệu như là phân bố dữ liệu xem quảng cáo qua các thông tin như giới tính, tuổi tác,...

2. Hướng phát triển

- Tìm hiểu các phương pháp tiền xử lý dữ liệu và trực quan hoá dữ liệu nâng cao hơn
- Bài báo cáo chỉ tập trung vào phân tích và tiền xử lý, chưa có thông tin về đưa các mô hình học máy để dự đoán. Vì thế nên xem xét để đưa thêm các phương pháp học máy vào.