Table of Contents

1. ASK 2
Metadata 2
Context 2
Requirements 2
2. PREPARE 2
3. PROCESS 4
4. ANALYZE 6
What is the impact of our website traffic on revenue? 6
Which products get us pageviews and revenue? 11
What customer segments are there? 14
EDA 14
Data Preprocessing 16
K-Means algorithm 17

Email: namthanhpham111@gmail.com

1. ASK

Metadata

Source	Description	Link
Customers & Transactions	Thông tin khách hàng và transaction, cả online lẫn instore	Customers
Website Traffic	ebsite traffic data cho từng trang sản phẩm (Từ tháng 1 đến tháng 12 năm 2020)	Traffic

Context

- Company A sells fashion. They have stores in: UK (London); FR (Paris); IT (Milan); GER (Berlin) - online or store.
- Thay have a webisite online store. In some cases, the page might have gone up after the product was put up for sale.

Requirements

- What is the impact of our website traffic on revenue?
- Which products get us pageviews and revenue?
- What customer segments are there?

2. PREPARE

- Bộ dữ liệu gồm 1 file Excel **customers (UK)** với 3 sheet: Customer info, Items, Customer transactions và 12 file **Traffic** ;
- Tool: Python;

Customer info

Out[2]:

	ID	FirstName	LastName	Country	DateJoined	Gender	Birthday	Newsletter
0	0.0	V0.296680287495188	L0.104646531512644	FR - France	2015-12-18	NaN	1968-02-03	N
1	1.0	D0.793097101838541	Law0.141693355411763	GER - Germany	2015-12-21	M	2009-10-06	Y
2	2.0	Ker0.141418247925814	Ng0.753960335680345	FR - France	2015-12-22	F	1990-08-04	Y
3	3.0	Fik0.950054552966336	F0.590961171612745	UK - United Kingdom	2015-12-22	M	1974-07-24	N
4	4.0	Iona0.294287981536498	Ison0.826191754811968	IT - Italy	2015-12-22	M	1981-08-13	N

Items

Email: namthanhpham111@gmail.com

Out[4]:

	ItemID	Product	Brand	SellPrice	CostPrice
0	1.0	032irview0.686128260621012	Ki)D3jDmA,RIP68X	943.0	359.0
1	2.0	070ttream0.518887735674677	GO4582ey<S!+k1VE	717.0	207.0
2	3.0	070htream0.333307794468401	G.Kb^jz*soY!(-4Q	739.0	199.0
3	4.0	100Grseys0.271522111052549	Dr|vm[-5p~56Y\mk	532.0	262.0
4	5.0	100[nside0.645837365801341	Dr|vm[-5p~56Y\mk	593.0	392.0

Customer transactions

Out[6]:

	OrderID	CustomerID	ItemID	TransactionDate	Channel
0	0.0	0.0	352.0	2020-03-21	In Store
1	0.0	0.0	3433.0	2020-07-14	In Store
2	0.0	0.0	11162.0	2020-08-09	In Store
3	0.0	0.0	13011.0	2020-12-07	In Store
4	0.0	0.0	13885.0	2020-11-08	In Store

- Gộp dữ liệu các file traffic:

Out[9]:

	Page URL	users	uniquePageviews	pageviews	Brand	Posted On (DD/MM/YYYY)
0	/2020/1/032irview0.686128260621012	5669.2	5777.8	6286.4	Ki)D3jDmA,RIP68X	2020-01-10 16:56:13
1	/2020/1/070ttream0.518887735674677	359.8	370.4	403.4	GO4582ey<S!+k1VE	2020-01-10 05:04:35
2	/2020/1/070htream0.333307794468401	587.6	614.2	657.6	G.Kb^jz*soY!(-4Q	2020-01-16 23:27:08
3	/2020/1/100Grseys0.271522111052549	1284.0	1308.6	1385.4	Dr|vm[-5p~56Y\mk	2020-01-17 12:32:24
4	/2020/1/100[nside0.645837365801341	1846.0	1880.8	2025.0	Dr|vm[-5p~56Y\mk	2020-01-23 05:21:08
...	...	...	...	...	...	...
17884	/2020/12/yinfbowls0.6558670149224	130.0	133.2	142.2	HoXbja)_x007f_qT:ESE8#	2020-12-01 04:55:35
17885	/2020/12/yosolease0.0455049365834417	314.6	325.6	365.2	YoRQW7@*&5W+~4Y"	2020-12-03 15:25:00
17886	/2020/12/you^-info0.127669879156147	1419.6	1448.4	1615.8	MouDd/fn,XEARhBp	2020-12-06 23:57:54
17887	/2020/12/youe-info0.729842478879044	483.8	487.6	509.6	yo4)AUKGso=,?!DX	2020-12-04 05:59:22
17888	/2020/12/zahJuseum0.120647638596675	823.2	837.6	886.8	Ch_x007f_w1fD(grL&W)5k	2020-12-05 23:49:29

17889 rows × 6 columns

- Tổng hợp dữ liệu để phân tích

Out[12]:

	OrderID	CustomerID	ItemID	TransactionDate	Channel	FirstName	LastName	Country	DateJoined	Gender	Birthday	Newsletter	
0	0	0	352	2020-03-21	In Store	V0.296680287495188	L0.104646531512644	FR - France	2015-12-18	NaN	1968-02-03	N	chrUrumorC
1	637	57	352	2020-03-15	In Store	Var0.0876970591979241	Bhan0.0592432181453122	GER - Germany	2016-02-02	F	1966-01-14	N	chrUrumorC
2	0	0	3433	2020-07-14	In Store	V0.296680287495188	L0.104646531512644	FR - France	2015-12-18	NaN	1968-02-03	N	corGvideoC
3	1840	167	3433	2020-05-16	In Store	Y0.694526245425698	Pur0.889218979929135	UK - United Kingdom	2015-04-16	M	1972-03-23	Y	corGvideoC
4	3386	307	3433	2020-04-23	In Store	Moh0.478682882191102	Khair0.475794987205303	FR - France	2016-06-02	NaN	1989-05-04	Y	corGvideoC
...	...	...	...	...	...	...	...	...	...	...	...	...	

Email: namthanhpham111@gmail.com

In [13]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 25213 entries, 0 to 25212
Data columns (total 16 columns):
 #   Column           Non-Null Count  Dtype         
---  ------           --------------  -----         
 0   OrderID          25213 non-null  int64         
 1   CustomerID       25213 non-null  int64         
 2   ItemID           25213 non-null  int64         
 3   TransactionDate  25213 non-null  datetime64[ns]
 4   Channel          25213 non-null  object        
 5   FirstName        25213 non-null  object        
 6   LastName         25213 non-null  object        
 7   Country          25213 non-null  object        
 8   DateJoined       25213 non-null  datetime64[ns]
 9   Gender           16892 non-null  object        
 10  Birthday         25213 non-null  datetime64[ns]
 11  Newsletter       25213 non-null  object        
 12  Product          25213 non-null  object        
 13  Brand            25213 non-null  object        
 14  SellPrice        25213 non-null  int64         
 15  CostPrice        25213 non-null  int64         
dtypes: datetime64[ns](3), int64(5), object(8)
memory usage: 3.3+ MB
```

3. PROCESS

Ta tiến hành gộp và xử lý định dạng của dữ liệu để tiến hành trả lời các câu hỏi

In [15]:
```
Channel_amounts
```

Out[15]:

TransactionDate	Channel	Orders	Total_items	Revenue
2020-01-01	In Store	30	30	23820
2020-01-01	Online	35	36	32044
2020-01-02	In Store	33	33	24069
2020-01-02	Online	40	40	31612
2020-01-03	In Store	46	46	37118
...	...	...	...	...

Email: namthanhpham111@gmail.com

In [28]:
```
# Đổi tên
traffic_rev.rename(columns = {'Date':'TransactionDate'}, inplace = True)
traffic_rev
```

Out[28]:

	TransactionDate	users	uniquePageviews	pageviews	Revenue
0	2020-01-01	63755.4	64814.8	70353.6	55864
1	2020-01-02	111861.4	113640.0	122557.8	55681
2	2020-01-03	63749.8	65175.4	71058.0	67753
3	2020-01-04	13946.0	14274.6	15349.6	64573
4	2020-01-05	47065.0	47856.6	51438.2	61098

In [42]:
```
#Tạo bảng tổng hợp doanh thu sản phẩm
item_revenue = data.groupby(['Brand', 'Product']).agg(TotalRevenue = ('SellPrice','sum'), TotalOrder = ('OrderID', 'count'), \
                                                      TotalCustomers = ('CustomerID', 'count')).reset_index()
```

In [43]:
```
#Lọc bảng
item_revenue.sort_values('TotalRevenue', ascending = False)
```

Out[43]:

	Brand	Product	TotalRevenue	TotalOrder	TotalCustomers
4885	II^(xPdB:S`#irqz	hypsrview0.321288570724117	11556	9	9
9876	Relp\+KJ?D,cWw0P	reeelease0.452821711209563	9205	7	7
306	Ac8lJsKH,4xtY.Tk	audE-info0.884915261087885	8428	7	7
3195	DiCo4(99zZ<nkafj	kinv-news0.29905739542661	7980	6	6
12575	YMbpE$ev3qMx-h*E	ymc_orson0.319251813809483	7794	6	6
...	...	...	...	...	...

In [58]:
```
# Lấy ra các cột cần tính toán
traff_rev = traff_rev[['Product', 'Brand', 'pageviews', 'TotalRevenue', 'traffic rank', 'ranking']]
traff_rev
```

Out[58]:

	Product	Brand	pageviews	TotalRevenue	traffic rank	ranking
0	101Dlease0.745305177696334	SSJ%#@$7LTf<p'Jx	305.6	465	Low	Low
1	a-qXailer0.39660536369098	MouDd/fn,XEARhBp	3705.8	444	Popular	Low
2	baiXpsule0.236876019278184	PaasB'_LIDEk`=W	726.8	3540	High	Very High
3	blopvideo0.67317091233294	VipZx1>S^s?}%X{}	38.0	1800	Low	High
4	fujBdence0.385183712192661	ToOYo&o|?(;MH>V:	4457.2	1689	Popular	High
...	...	...	...	...	...	...
12365	trobhotos0.430881773750675	Jo?$J}r__x007f_"0#/!m:	4030.2	3699	Popular	Very High
12366	twiO-word0.109741844099603	Mc&v?E*9%}~WP@rD	889.8	1536	High	High
12367	vanAction0.315573410341238	Vag9"-Z=gN30ND@[	330.4	1162	Low	Medium
12368	vanMtists0.0499183505700914	Vag9"-Z=gN30ND@[	25109.6	1208	Popular	High
12369	vetssalia0.105184586361579	Relp\+KJ?D,cWw0P	305.4	329	Low	Low

12370 rows × 6 columns

Email: namthanhpham111@gmail.com

In [64]:
```
customer_data = data.groupby(["CustomerID","FirstName", "LastName","Country","Birthday","DateJoined","Newsletter"]).\
                        agg(total_expenditures = ('SellPrice','sum')).\
                        reset_index()
customer_data.head()
```

Out[64]:

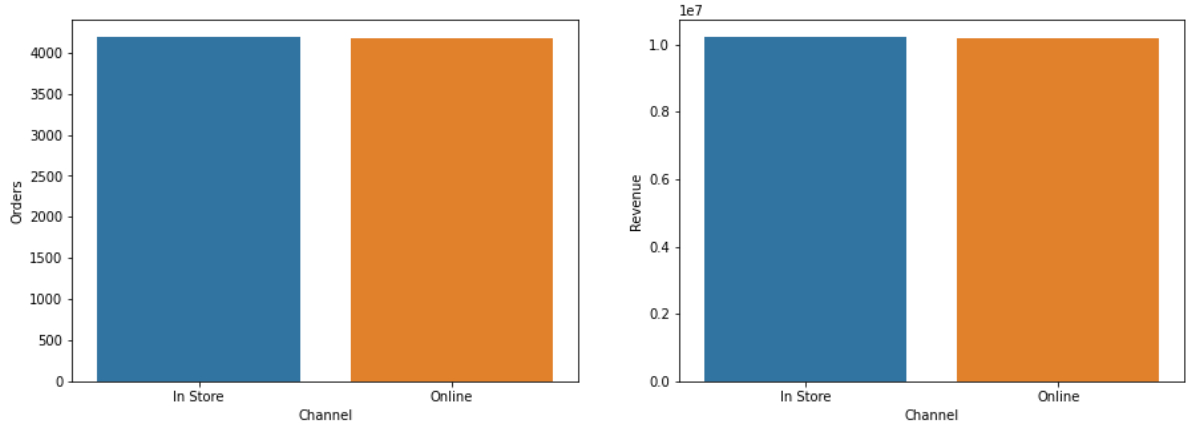
	CustomerID	FirstName	LastName	Country	Birthday	DateJoined	Newsletter	total_expenditures
0	0	V0.296680287495188	L0.104646531512644	FR - France	1968-02-03	2015-12-18	N	32997
1	1	D0.793097101838541	Law0.141693355411763	GER - Germany	2009-10-06	2015-12-21	Y	34948
2	2	Ker0.141418247925814	Ng0.753960335680345	FR - France	1990-08-04	2015-12-22	Y	17879
3	3	Fik0.950054552966336	F0.590961171612745	UK - United Kingdom	1974-07-24	2015-12-22	N	14603
4	4	Iona0.294287981536498	Ison0.826191754811968	IT - Italy	1981-08-13	2015-12-22	N	37029

4. ANALYZE

Trả lời 3 câu hỏi của bài test

What is the impact of our website traffic on revenue?

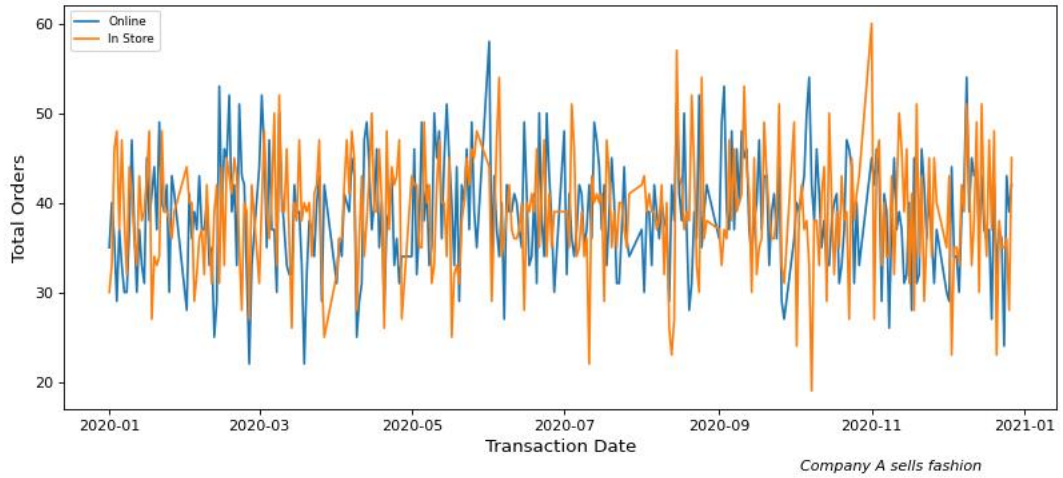
	Channel	Orders	Revenue
0	In Store	4192	10222263
1	Online	4171	10194955



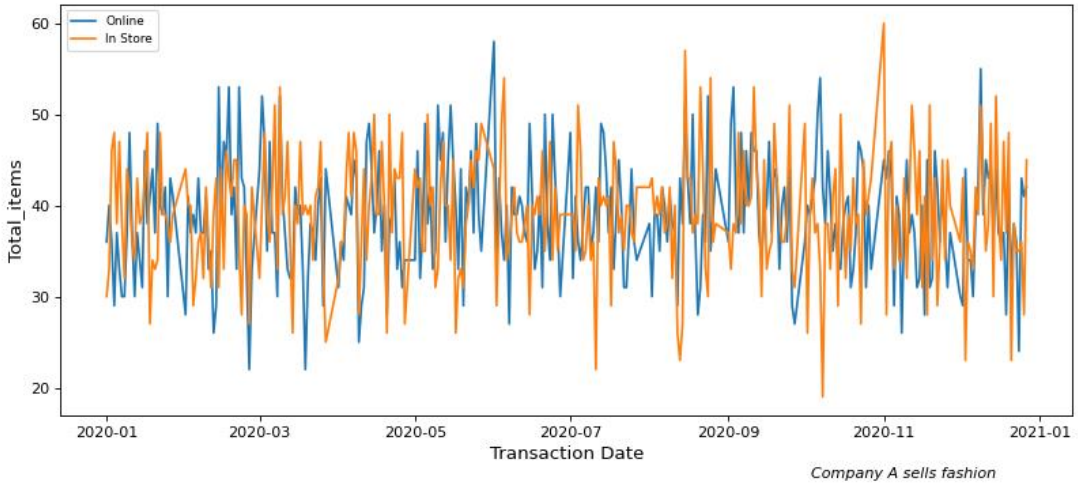
Biểu đồ thể hiện Orders và Revenue theo Transaction date của 2 kênh bán hàng là Online và In Store có tỉ lệ tương đương nhau.

Email: namthanhpham111@gmail.com

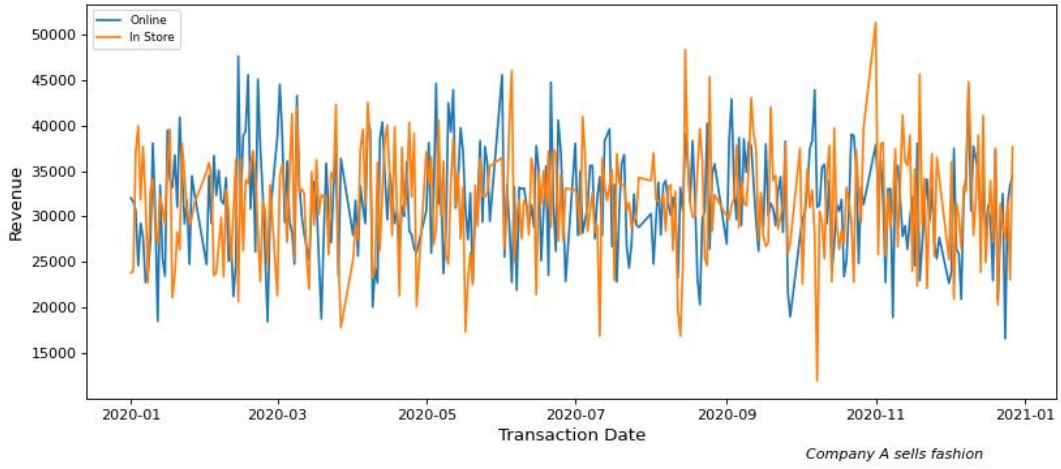
Total Orders by Date



Total Items by Date



Revenue by Date

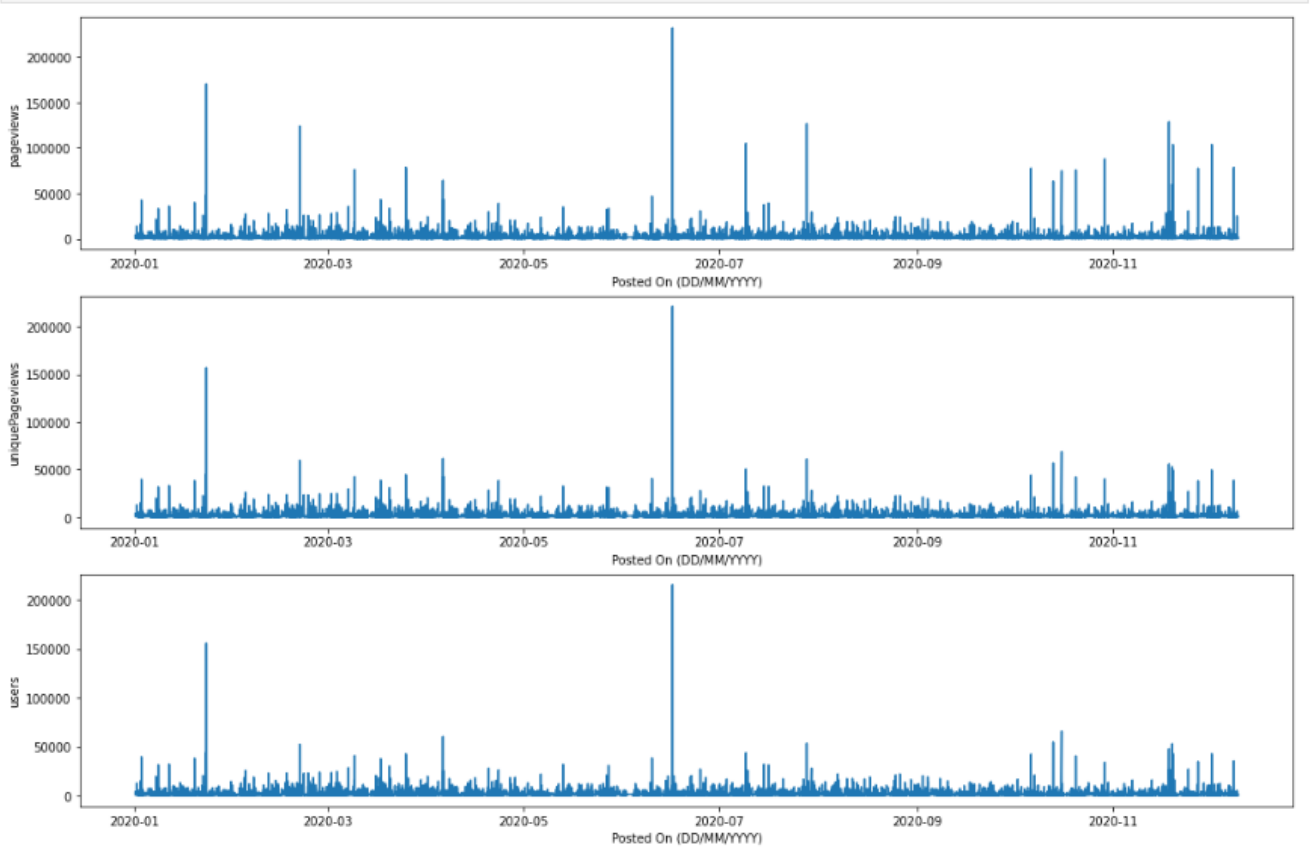


Email: namthanhpham111@gmail.com

Tổng kết

---> Tổng số orders, items và revenue của 2 kênh Online và In Store có diễn biến theo thời gian tương đương nhau.

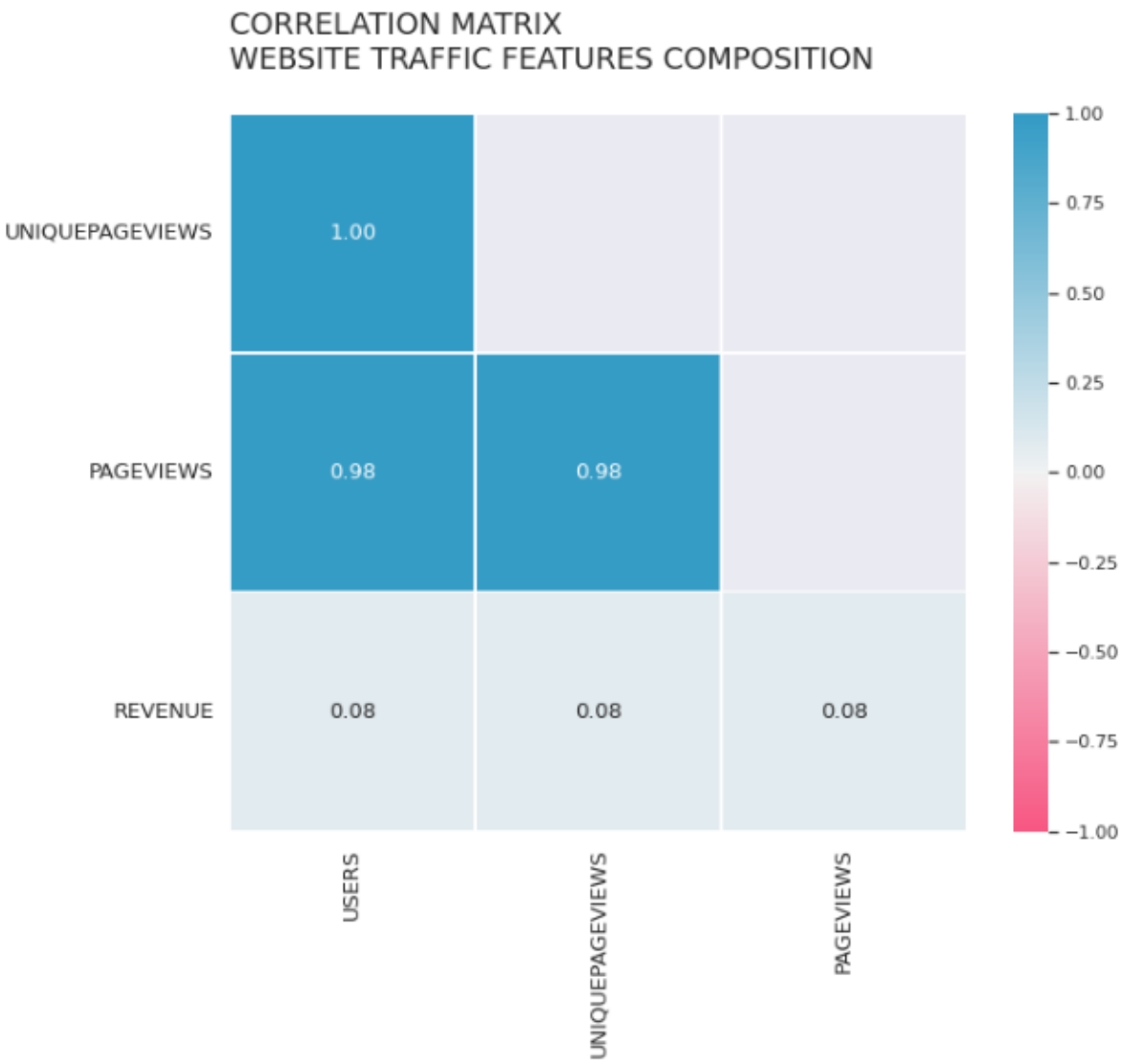
- Users, uniquePageviews, pageviews by Date



- Total revenue vs total users, uniquePageviews, pageviews

Email: namthanhpham111@gmail.com

Out[35]: Text(0.0, 1.0, 'CORRELATION MATRIX\nWEBSITE TRAFFIC FEATURES COMPOSITION\n')



Email: namthanhpham111@gmail.com

P-value

P-value là giá trị xác suất mà mối tương quan giữa hai biến này có ý nghĩa thống kê. Thông thường, chúng ta chọn mức ý nghĩa 0,05, nghĩa là chúng ta tin tưởng 95% rằng mối tương quan giữa các biến là có ý nghĩa.

Với các giá trị:

- the p-value > 0.001: Mối tương quan giữa 2 biến mạnh
- the p-value > 0.05: Mối tương quan giữa 2 biến tốt
- the p-value >= 0.1: Tương quan yếu
- the p-value < 0.1: there is no evidence that the correlation is significant.

Email: namthanhpham111@gmail.com

In [38]:
```
# Mối quan hệ giữa Revenue và user
pearson_coef, p_value = stats.pearsonr(traffic_rev['Revenue'], traffic_rev['users'])
print("Correlation Coefficient is", pearson_coef.round(3))
print("Values P-value is ", p_value.round(3))
```

```
Correlation Coefficient is 0.076
Values P-value is  0.184
```

In [39]:
```
# Mối quan hệ giữa Revenue và pageviews
pearson_coef, p_value = stats.pearsonr(traffic_rev['Revenue'], traffic_rev['pageviews'])
print("Correlation Coefficient is", pearson_coef.round(3))
print("Values P-value is ", p_value.round(3))
```

```
Correlation Coefficient is 0.075
Values P-value is  0.189
```

In [40]:
```
# Mối quan hệ giữa Revenue và uniquePageviews
pearson_coef, p_value = stats.pearsonr(traffic_rev['Revenue'], traffic_rev['uniquePageviews'])
print("Correlation Coefficient is", pearson_coef.round(3))
print("Values P-value is ", p_value.round(3))
```

```
Correlation Coefficient is 0.076
Values P-value is  0.188
```

Trả lời: Qua các biểu đồ trên và chỉ số P-value ta có thể thấy các Website traffic không ảnh hưởng đến doanh thu của Company A do có sự tương quan thấp.

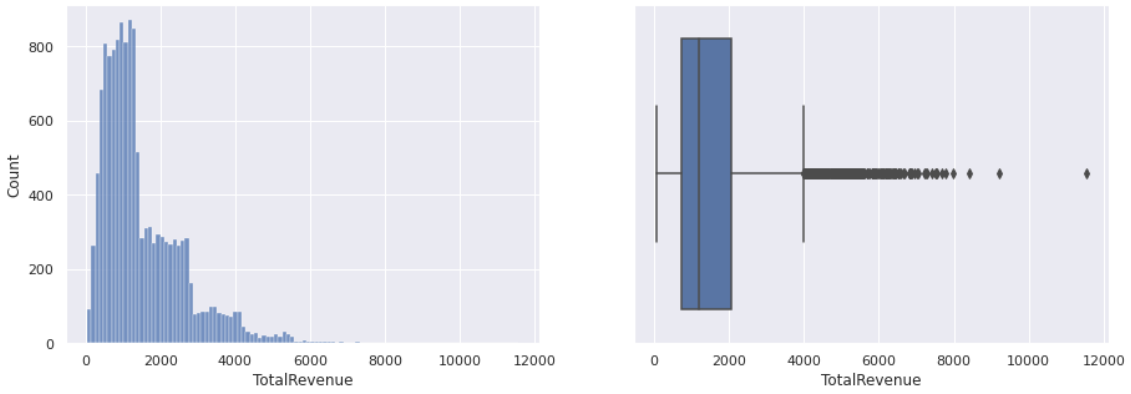
Which products get us pageviews and revenue?

- Sự phân bố của tổng doanh thu

In [45]:
```
# Sự phân bố của TotalRevenue
fig, axes = plt.subplots(1, 2, figsize=(15,5))

sns.histplot(ax=axes[0], x = 'TotalRevenue', data = item_revenue)
sns.boxplot(ax=axes[1], x = 'TotalRevenue', data = item_revenue)
```

Out[45]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fc8d45e5850>
```



Email: namthanhpham111@gmail.com

- Top 10 sản phẩm có revenue cao nhất

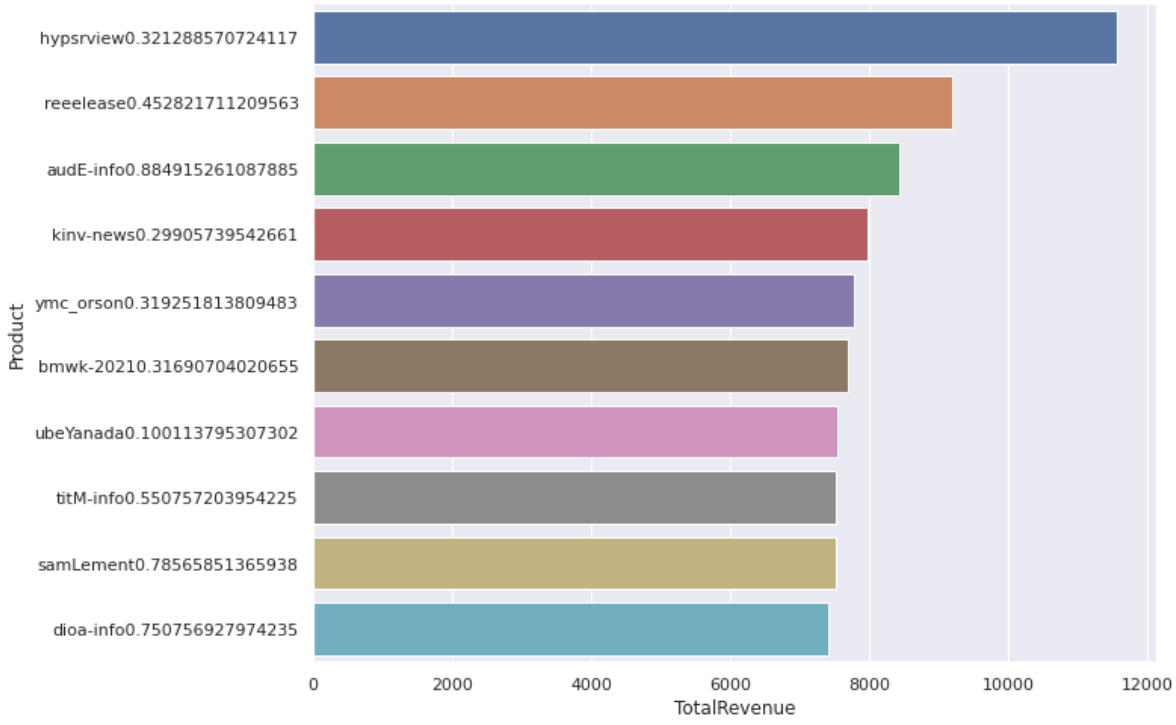
Top 10 sản phẩm có revenue cao nhất

In [47]:
```
# Top 10 sản phẩm có revenue cao nhất
item_revenue_product_10  = item_revenue[item_revenue['ranking'] == 'Very High'].sort_values('TotalRevenue', ascending = False).head(10)

item_revenue_product_10
```

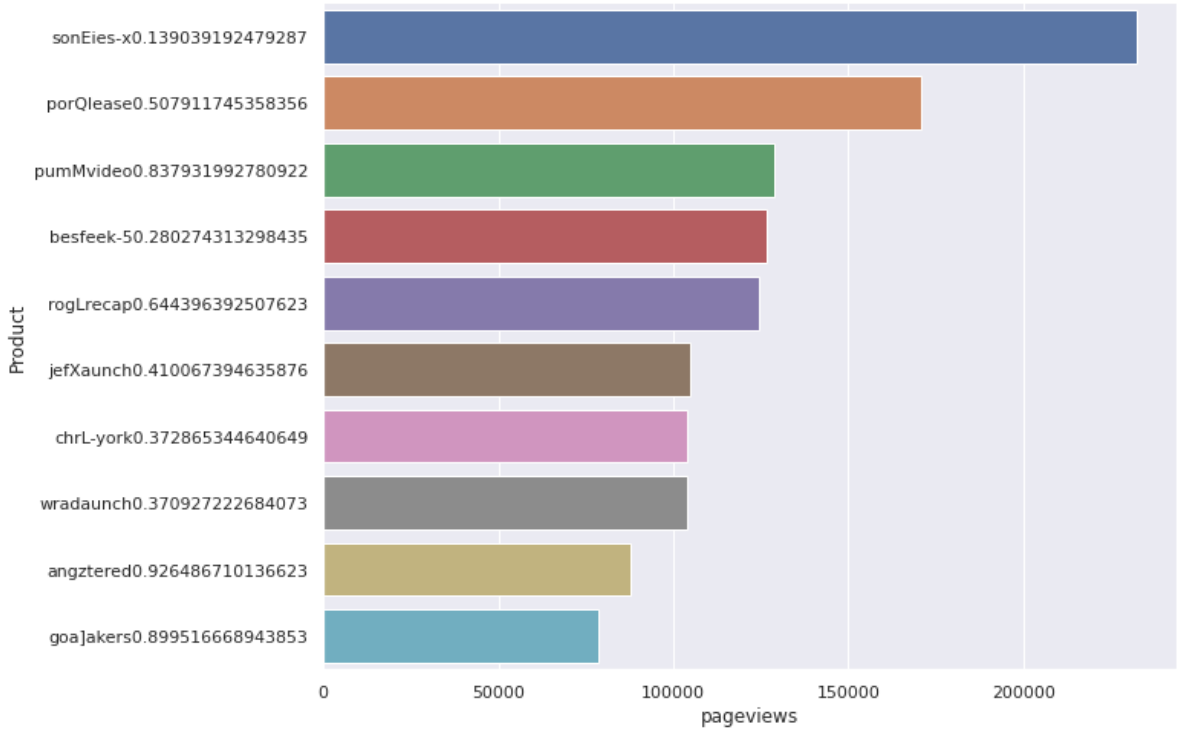
Out[47]:

	Brand	Product	TotalRevenue	TotalOrder	TotalCustomers	ranking
4885	II^(xPdB:S`#irqz	hypsrview0.321288570724117	11556	9	9	Very High
9876	Relp\+KJ?D,cWw0P	reeelease0.452821711209563	9205	7	7	Very High
306	Ac8IJsKH,4xtY.Tk	audE-info0.884915261087885	8428	7	7	Very High
3195	DiCo4(99zZ<nkafj	kinv-news0.29905739542661	7980	6	6	Very High
12575	YMbpE$ev3qMx-h*E	ymc_orson0.319251813809483	7794	6	6	Very High
1693	BMh._Fx~"+dbZy|	bmwk-20210.31690704020655	7686	6	6	Very High
1407	Ap8r`F[]sfP_x007f_->SZ	ubeYanada0.100113795307302	7552	4	4	Very High
13152	go@c<p.bPWb1nLrF	titM-info0.550757203954225	7520	5	5	Very High
10155	Sa?9zXUH5iJbuE'S	samLement0.78565851365938	7518	6	6	Very High
3231	DiL%1<G?YSFM_/TR	dioa-info0.750756927974235	7422	6	6	Very High



- Top 10 sản phẩm có traffic cao nhất

Email: namthanhpham111@gmail.com



Nhận xét: Những sản phẩm có revenue cao nhất thì lượng traffic không quá cao.

- Top 5 sản phẩm vừa có traffic cao vừa có revenue cao;

```
In [60]: # Top 5 sản phẩm vừa có traffic cao vừa có revenue cao
traff_rev[(traff_rev['ranking'] == 'Very High') & (traff_rev['traffic rank'] == 'Popular')].sort_values(['Total
```

Out[60]:

	Product	Brand	pageviews	TotalRevenue	traffic rank	ranking
9714	samLement0.78565851365938	Sa?9zXUH5iJbuE'S	1590.0	7518	Popular	Very High
6946	takgeveal0.557457939403364	Ka>n{\].q.~P&*Ao	8596.2	7049	Popular	Very High
3404	stupponyo0.211329698054268	MouDd/fn,XEARhBp	7190.6	6954	Popular	Very High
11	poryn-9920.309751616966229	PoAqnL=>P9Qb"ZUa	10795.6	6815	Popular	Very High
11075	bbcF-date0.977769785530841	Relp\+KJ?D,cWw0P	2049.6	6815	Popular	Very High

- Top 5 sản phẩm vừa có traffic Low nhưng có revenue cao;

Out[61]:

	Product	Brand	pageviews	TotalRevenue	traffic rank	ranking
2147	reeelease0.452821711209563	Relp\+KJ?D,cWw0P	281.8	9205	Low	Very High
9385	ymc_orson0.319251813809483	YMbpE$ev3qMx-h*E	223.2	7794	Low	Very High
9535	ausDition0.316805159126848	Exa~e4sZi*|fpD<^	161.4	7232	Low	Very High
4985	theftream0.650769488105747	Mi+QB'FqF;cGT)Y'	350.2	7000	Low	Very High
8121	raejement0.618482160856154	Ra&3X!d15ID^=sus	342.4	6860	Low	Very High

Email: namthanhpham111@gmail.com

- Top 5 sản phẩm vừa có traffic cao nhưng có revenue thấp;

Out[62]:

	Product	Brand	pageviews	TotalRevenue	traffic rank	ranking
8206	chaPtopia0.305759133071658	Tr1)L/A]=of['Qhn	1906.0	741	Popular	Low
4933	whil-info0.511259246630254	ad(;%f6iD'}9EHD[	1737.6	741	Popular	Low
9727	youTcture0.944314892191284	Ap!FulqmT[82a2/E	3566.4	740	Popular	Low
1874	niku5-3000.815662316070307	NiPea$n|Eu@<@>'L	2231.6	740	Popular	Low
3592	nik\-info0.17636213285294	NiPea$n|Eu@<@>'L	2175.4	740	Popular	Low

What customer segments are there?

- Tổng hợp dữ liệu

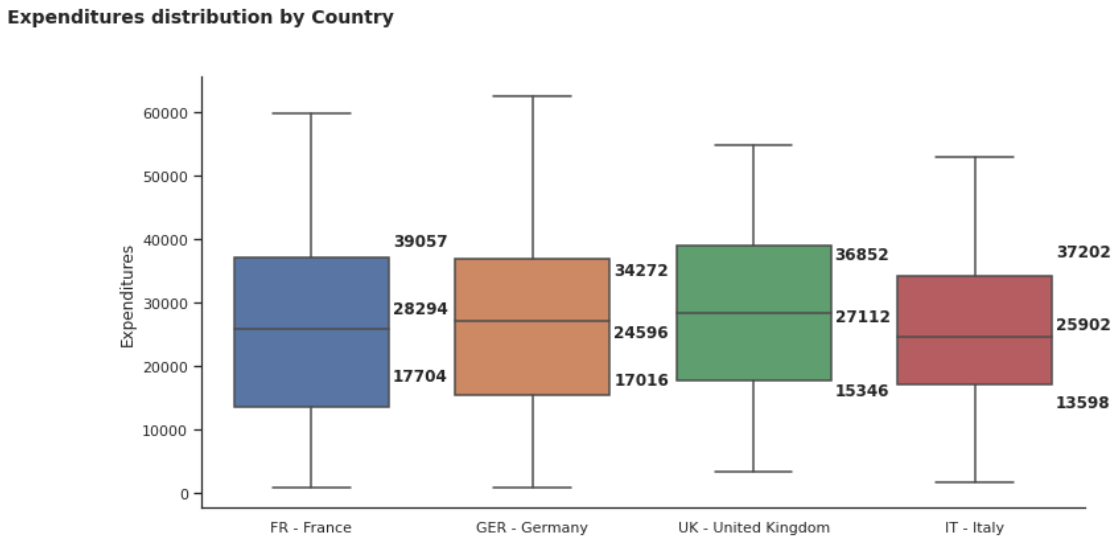
In [67]:

```
#merge với bảng customer
customer_seg = customer_data.merge(customer_info[['ID','Gender']],how='inner',left_on='CustomerID', right_on='ID')
customer_seg.drop('ID', inplace=True, axis=1) # bỏ 1 cột id
customer_seg.head()
```

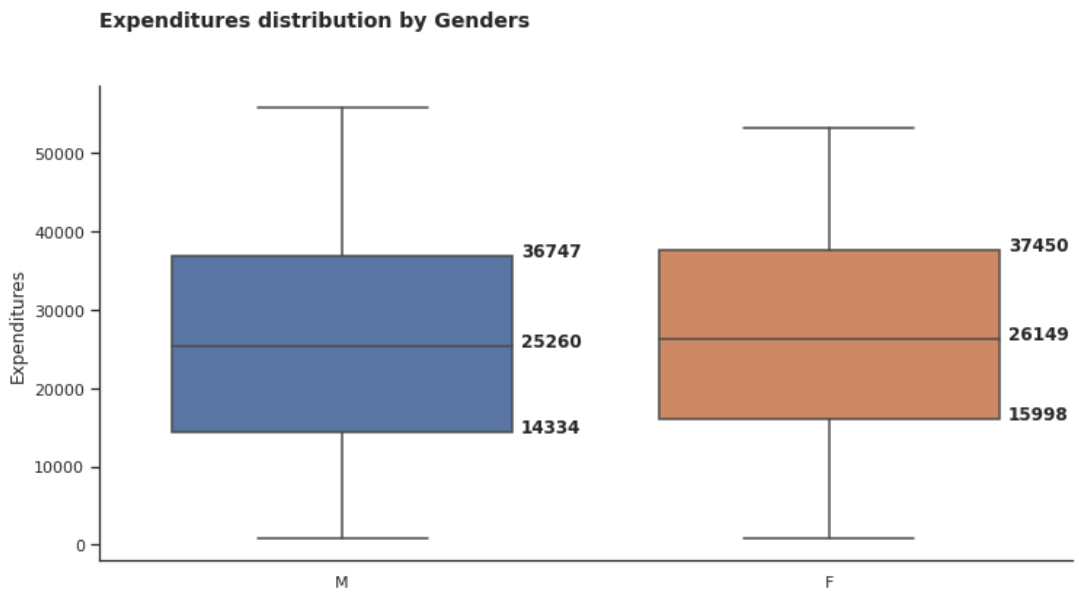
Out[67]:

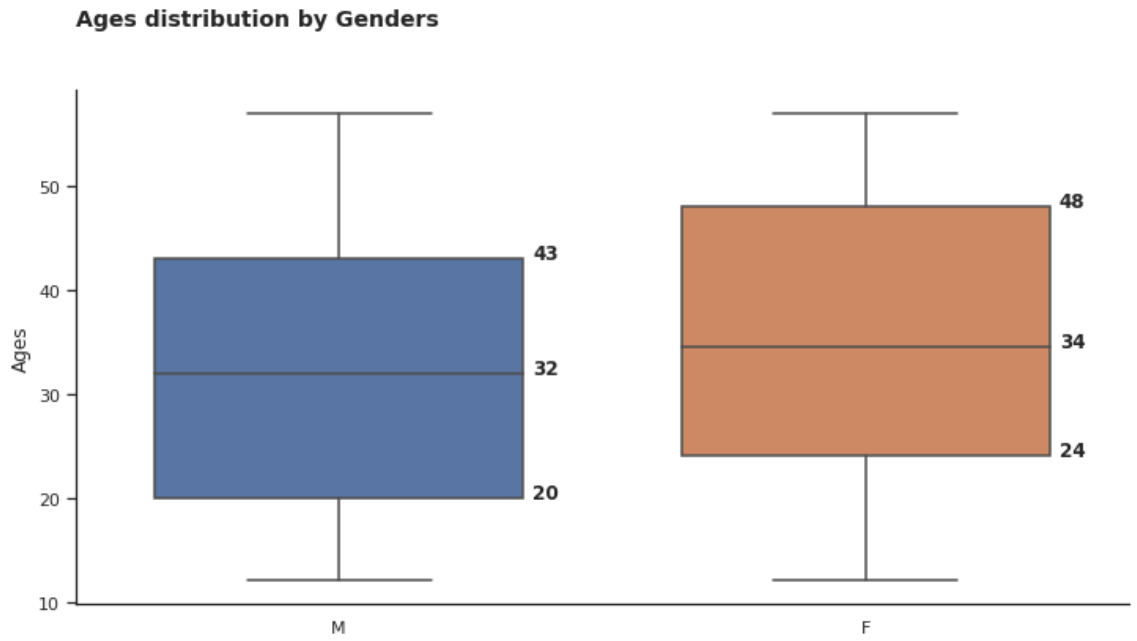
	CustomerID	FirstName	LastName	Country	Birthday	DateJoined	Newsletter	total_expenditures	Age	Loyalty	Gender
0	0	V0.296680287495188	L0.104646531512644	FR - France	1968-02-03	2015-12-18	N	32997	54	7	NaN
1	1	D0.793097101838541	Law0.141693355411763	GER - Germany	2009-10-06	2015-12-21	Y	34948	13	7	M
2	2	Ker0.141418247925814	Ng0.753960335680345	FR - France	1990-08-04	2015-12-22	Y	17879	32	7	F
3	3	Fik0.950054552966336	F0.590961171612745	UK - United Kingdom	1974-07-24	2015-12-22	N	14603	48	7	M
4	4	Iona0.294287981536498	Ison0.826191754811968	IT - Italy	1981-08-13	2015-12-22	N	37029	41	7	M

EDA



Email: namthanhpham111@gmail.com





Kết luận

- Không có sự khác biệt nhiều về mức chi tiêu giữa các khách hàng trung thành và không có outlier.
- Cả 4 thủ đô đều có mức chi tiêu từ hơn 10.000 (USD) đến gần 40.000 (USD). và không có sự khác biệt nhiều về mức chi tiêu cũng như không có outlier.

Email: namthanhpham111@gmail.com

Data Preprocessing

Label Encoding

Chuyển các feature dạng catergory thành dạng number để đưa vào mô hình

In [72]:
```
# tạo dữ liệu mới
segment = customer_seg.copy()
segment.head()
```

Out[72]:

	CustomerID	FirstName	LastName	Country	Birthday	DateJoined	Newsletter	total_expenditures	Age	Loyalty	Gender
0	0	V0.296680287495188	L0.104646531512644	FR - France	1968-02-03	2015-12-18	N	32997	54	7	NaN
1	1	D0.793097101838541	Law0.141693355411763	GER - Germany	2009-10-06	2015-12-21	Y	34948	13	7	M
2	2	Ker0.141418247925814	Ng0.753960335680345	FR - France	1990-08-04	2015-12-22	Y	17879	32	7	F
3	3	Fik0.950054552966336	F0.590961171612745	UK - United Kingdom	1974-07-24	2015-12-22	N	14603	48	7	M
4	4	Iona0.294287981536498	Ison0.826191754811968	IT - Italy	1981-08-13	2015-12-22	N	37029	41	7	M

Out[74]:

	Birthday	Country	Newsletter	LastName	FirstName	DateJoined	Gender
0	49	0	0	232	506	31	2
1	741	1	1	242	44	32	1
2	408	0	1	308	155	33	0
3	141	3	0	137	75	33	1
4	245	2	0	184	104	33	1

Standardization

Scale dữ liệu để đưa vào mô hình

In [76]:
```
data_scaled.head()
```

Out[76]:

	total_expenditures	Age
0	32997	54
1	34948	13
2	17879	32
3	14603	48
4	37029	41

(770, 2)

	total_expenditures	Age
0	0.479662	1.544695
1	0.624054	-1.602718
2	-0.639206	-0.144161
3	-0.881660	1.084098
4	0.778067	0.546735

Email: namthanhpham111@gmail.com

Dimensionality Reduction

Giảm chiều dữ liệu, là sự biến đổi dữ liệu từ không gian chiều-cao thành không gian chiều-thấp để biểu diễn ở dạng chiều-thấp đồng thời giữ lại một số thuộc tính có ý nghĩa của dữ liệu gốc, có ý tưởng là gần với chiều nội tại.

Các feature vectors trong các bài toán thực tế có thể có số chiều rất lớn, tới vài nghìn. Ngoài ra, số lượng các điểm dữ liệu cũng thường rất lớn. Nếu thực hiện lưu trữ và tính toán trực tiếp trên dữ liệu có số chiều cao này thì sẽ gặp khó khăn cả về việc lưu trữ và tốc độ tính toán. Vì vậy, giảm số chiều dữ liệu là một bước quan trọng trong nhiều bài toán. Đây cũng được coi là một phương pháp nén dữ liệu.

Trong bài test ta sẽ sử dụng phương pháp đơn giản nhất trong các thuật toán Dimensionality Reduction dựa trên một mô hình tuyến tính. Phương pháp này có tên là Principal Component Analysis (PCA), tức Phân tích thành phần chính. Phương pháp này dựa trên quan sát rằng dữ liệu thường không phân bố ngẫu nhiên trong không gian mà thường phân bố gần các đường/mặt đặc biệt nào đó. PCA xem xét một trường hợp đặc biệt khi các mặt đặc biệt đó có dạng tuyến tính là các không gian con (subspace).

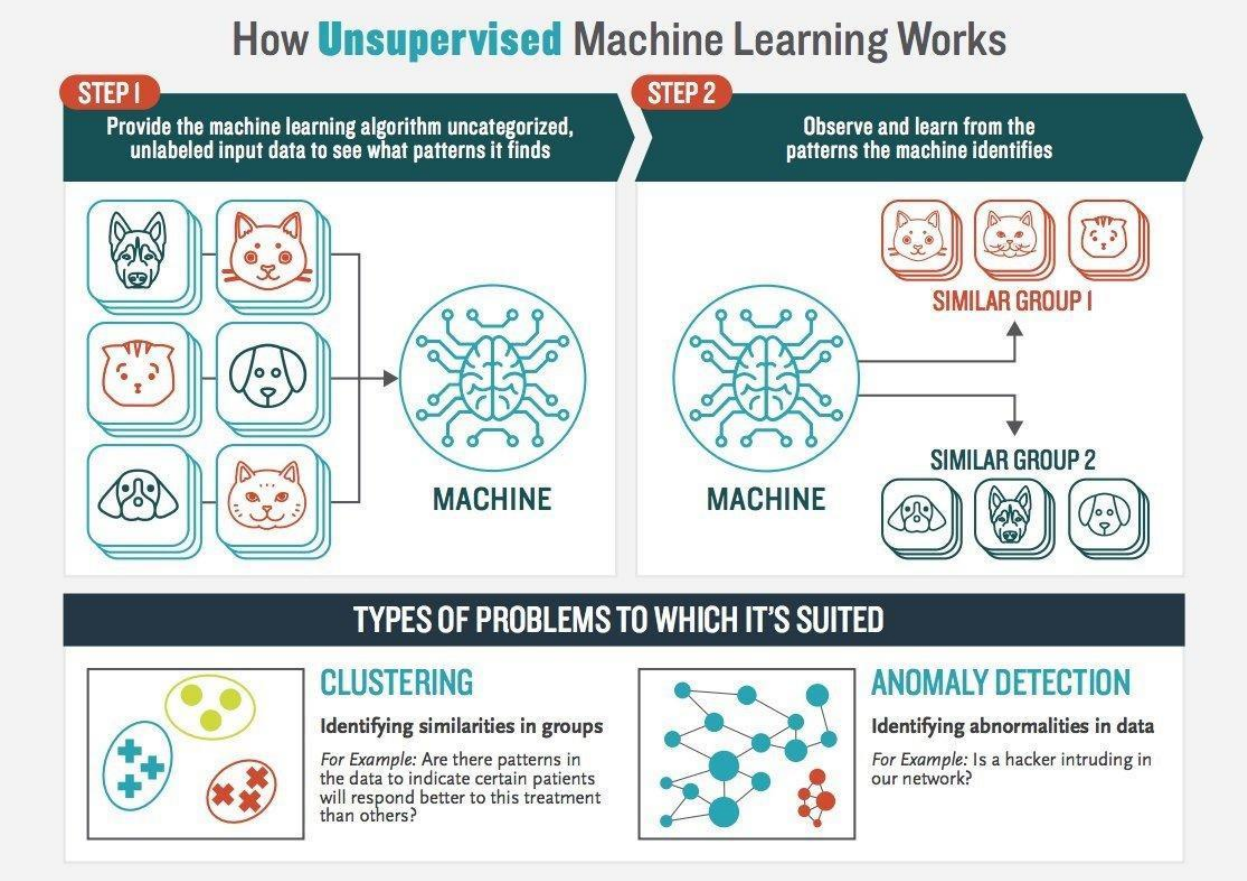
Out[84]:

	count	mean	std	min	25%	50%	75%	max
PC1	770.0	-2.306957e-18	1.004038	-2.427659	-0.707849	0.012207	0.715837	2.433561

K-Means algorithm

Trong bài test ta sẽ sử dụng thuật toán K – means để tìm ra các phân khúc khách hàng.

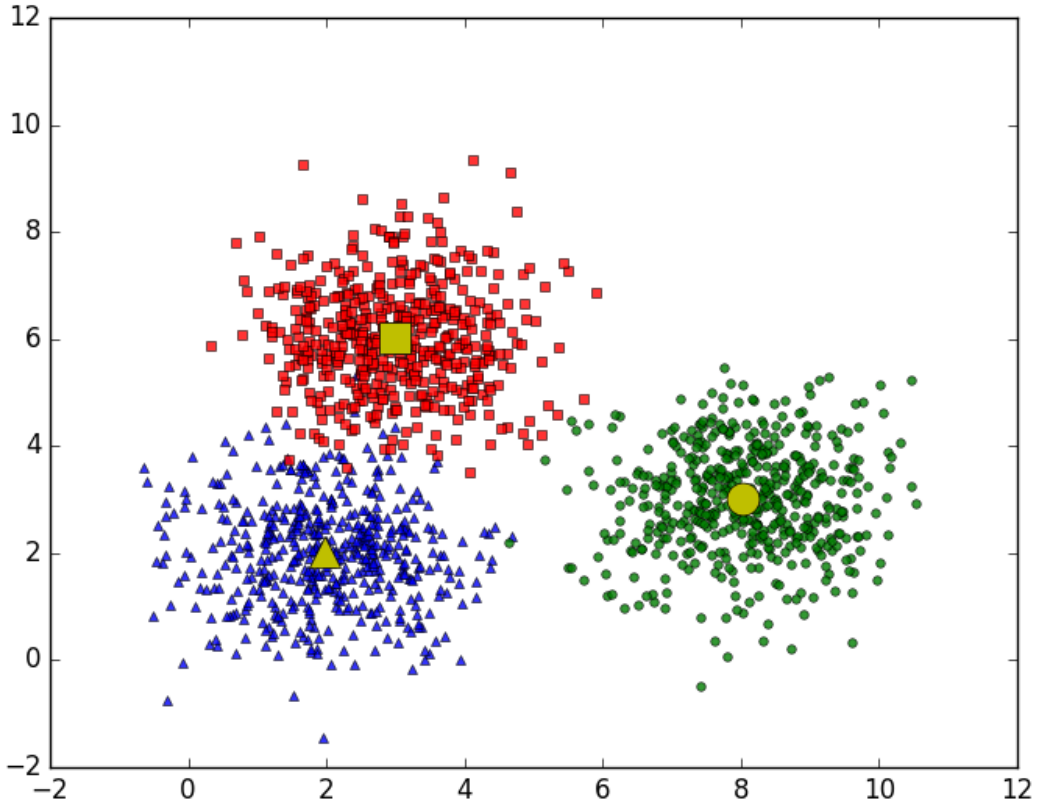
K-Means là một trong 3 thuật toán chính nhóm Clustering thuộc nhánh Unsupervised Learning của Machine Learning.



K Means Clustering hay phân cụm là phương pháp tập hợp các điểm ở gần nhau trong một không gian nào đó (không gian có thể là 2D , 3D thậm chí ND)

Email: namthanhpham111@gmail.com

Ví dụ: 3 cụm dữ liệu với dữ liệu 2D

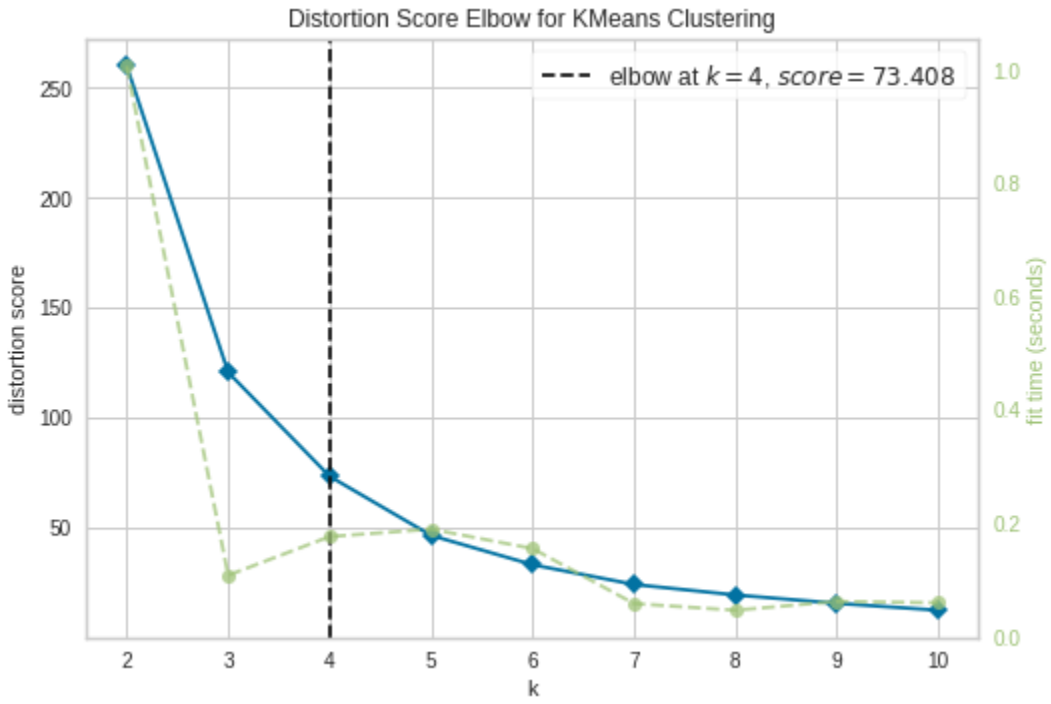


Và trong thuật toán Clustering , số K (tức là tâm cụm) ban đầu chúng ta khai báo rất quan trọng đối với kết quả đầu ra, vì nếu K quá nhiều hay quá ít thì cũng có thể làm cho kết quả đầu ra không có ý nghĩa (chúng ta hay hiểu là overfit và underfit)

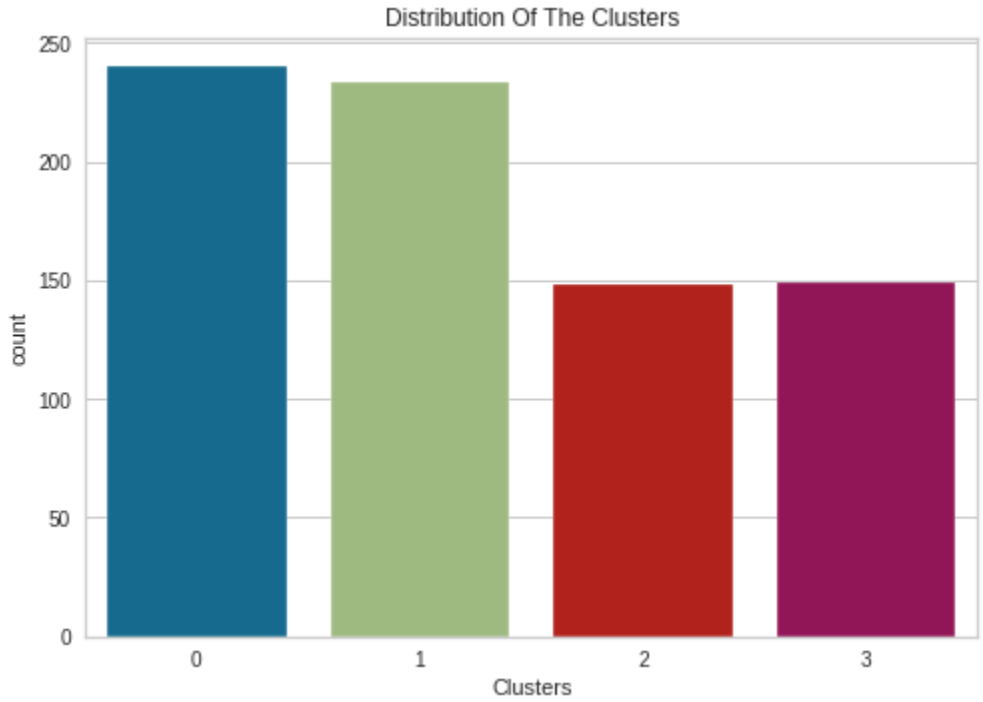
Đối với thuật toán Unsupervised , việc có dữ liệu đầu vào đồng đều , chọn thuật toán phù hợp và cơ số k phù hợp nó quan trọng hơn việc chúng ta adjust các chỉ số để cho ra output có ý nghĩa, vì vật , phần toán học của thuật toán như hàm tổng quan , hàm mất mát nó không quan trọng bằng việc chúng ta chọn được cơ số K phù hợp và thuật toán phù hợp.

Chạy mô hình

- Visual model K-means

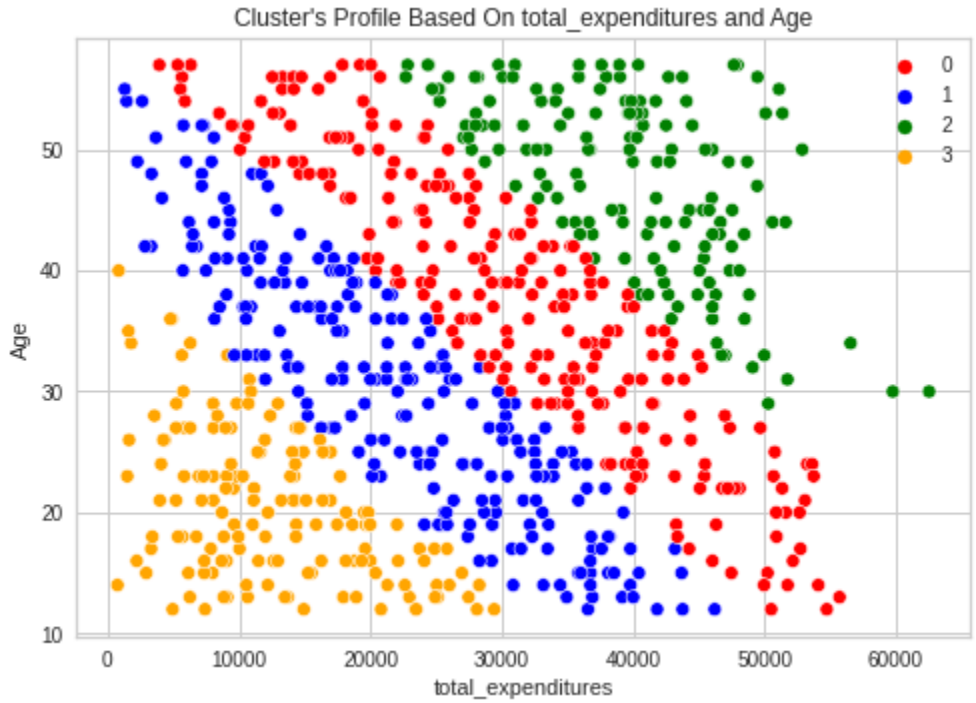


Ta sẽ chọn hệ số K là 4 dựa theo điểm phân bố



Email: namthanhpham111@gmail.com

- Visualize phân bố của mô hình K – Means



Kết luận

Phân khúc khúc khác hàng được chia thành 4 nhóm gồm:

- Nhóm 1 : Nhóm khách hàng ***từ 16 - 25 tuổi***, có mức chi tiêu từ ***7.500 - 18.000 USD*** (Cluster 3);
- Nhóm 2: Nhóm khách hàng từ ***22 - 38 tuổi***, có mức chi tiêu từ ***14.000 - 31.000 USD*** (Cluster 1);
- Nhóm 3: Nhóm khách hàng từ ***30 - 47 tuổi***, có mức chi tiêu ***từ 22.000 - 40.000 USD*** (Cluster 0);
- Nhóm 4: Nhóm khách hàng từ ***43 - 54 tuổi***, có mức chi tiêu từ ***34.000 - 46.000 USD*** (Cluster 2).