

ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ



Họ và tên: Nguyễn Tuấn Thành

Mã sinh viên: 22022624

Đề án môn: Lập trình xử lý dữ liệu với Python

Ngành: Trí tuệ nhân tạo

Cán bộ hướng dẫn: Tiến sĩ Đặng Trần Bình

Thạc sĩ Nguyễn Văn Phi

Thạc sĩ Đỗ Hải Sơn

Hà Nội – 2023

Nội dung

Trong thực tế hiện nay việc cào dữ liệu Facebook là một việc vô cùng cần thiết và có ứng dụng rất rộng rãi . Nhiều công ty về truyền thông ở Việt Nam đã làm nhiều năm nay đã thực hiện cào dữ liệu để thu thập thông tin và đo mức độ ảnh hưởng cho các chương trình marketing / quảng cáo. Một số công ty khác thì bán giải pháp trích xuất thông tin từ Facebook với đa dạng các nội dung có thể lấy được. Hoặc đơn giản nhất với sinh viên hoặc một số shop bán hàng nhỏ họ thực hiện cào dữ liệu page facebook để theo dõi tình trạng phát triển của page, xu hướng phát trong thời gian gần đây. Trong một số cuộc thi về truyền thông cần tìm tất cả các bài viết Facebook để tìm xem bài viết nào nhiều tương tác nhất để giao giải thưởng , hoặc trong 1 bài viết / livestream ta cần cào tất cả các bình luận để xem xem ai là người bình luận sớm nhất và đúng nhất ,....Việc cào dữ liệu Facebook là một ý tưởng rất hay và cần thiết để nghiên cứu . Chính vì vậy trong đề tài này em đã thực hiện cào dữ liệu Facebook về để nghiên cứu và phân tích dựa vào các trường thông tin đã thu thập được. Do chưa có nhiều kinh nghiệm về làm báo cáo cũng như những hạn chế về mặt kiến thức , trong bài báo cáo chắc chắn cũng không thể tránh khỏi những thiếu sót . Em mong nhận được sự đóng góp , phê bình từ phía thầy để báo cáo của em được hoàn thiện hơn .

Em xin chân thành cảm ơn thầy!

Hà Nội, Tháng 12 năm 2023

Thành

Nguyễn Tuấn Thành

Mục lục

Phần 1 : Thu thập dữ liệu:

1.1 Thư viện dùng	5
1.2 File www.facebook.cookies.txt	5
1.3 Thu thập dữ liệu về	6
1.4 Lưu dữ liệu về	7
1.5 Xem thông tin dữ liệu thu thập được	8

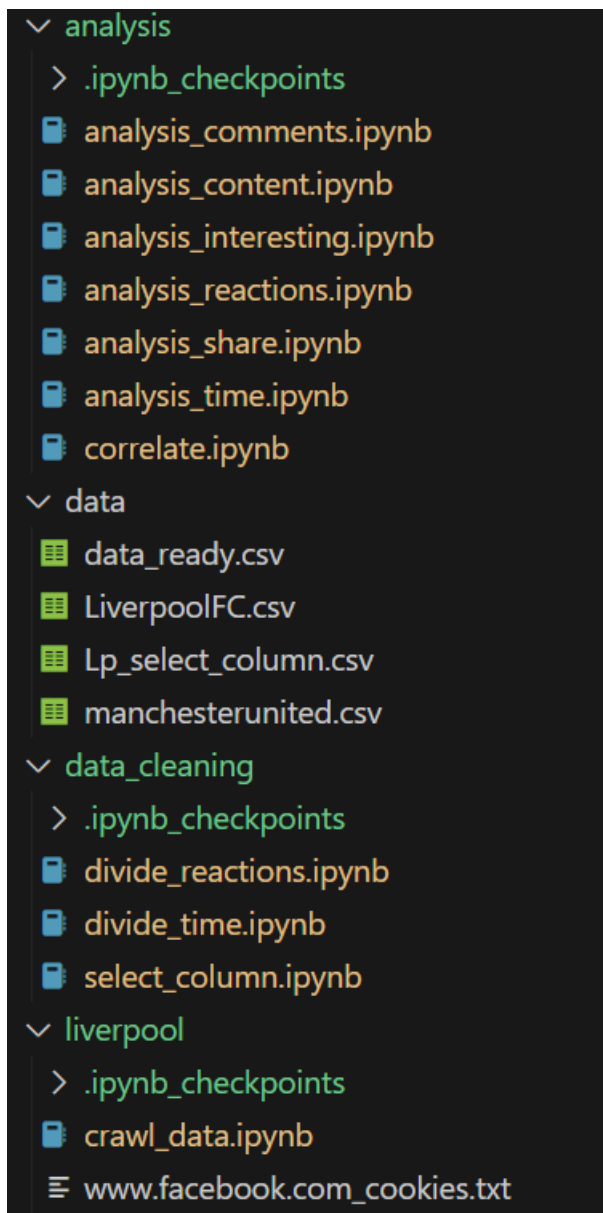
Phần 2 : Làm sạch và xử lí dữ liệu:

2.1 Chọn những cột cần dùng và loại bỏ dữ liệu đúng	9
2.2 Tách cột biểu cảm thành các trường cần phân tích	11
2.3 Tách cột thời gian thành các cột ngày tháng năm , thứ, giờ	13

Phần 3 : Phân tích dữ liệu:

3.1 Phân tích nội dung	17
3.2 Phân tích bình luận	19
3.3 Phân tích biểu cảm	22
3.4 Phân tích lượt chia sẻ	31
3.5 Phân tích thời gian đăng bài	37
3.6 Phân tích tương quan các trường dữ liệu	42
3.7 Phân tích thứ vị về dữ liệu	51

Các tệp có khung sau:



Link GitHub: https://github.com/TuanThanh2004/project_analysis_data_facebook.git

Phần 1 : Thu thập dữ liệu từ 2 trang Facebook và lưu dữ liệu:

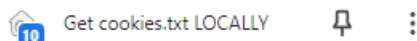
1.1 Thư viện dùng:

Thư viện thu thập dữ liệu thông tin các bài viết từ 1 trang facebook để người dùng có thể sử dụng để phân tích.

```
Entrée [1]: 1 %pip install facebook_scraper pandas numpy
Requirement already satisfied: dateparser<2.0.0,>=1.0.0 in c:\users\admin\anaconda3\lib\site-packages (from facebook_scraper) (1.1.8)
Requirement already satisfied: demjson3<4.0.0,>=3.0.5 in c:\users\admin\anaconda3\lib\site-packages (from facebook_scraper) (3.0.6)
Requirement already satisfied: requests-html<0.11.0,>=0.10.0 in c:\users\admin\anaconda3\lib\site-packages (from facebook_scraper) (0.10.0)
Requirement already satisfied: python-dateutil<=2.8.1 in c:\users\admin\anaconda3\lib\site-packages (from pandas) (2.8.2)
Requirement already satisfied: pytz<=2020.1 in c:\users\admin\anaconda3\lib\site-packages (from pandas) (2022.7)
Requirement already satisfied: regex!=2019.02.19,!=2021.8.27 in c:\users\admin\anaconda3\lib\site-packages (from dateparser<2.0.0,>=1.0.0->facebook_scraper) (2022.7.9)
Requirement already satisfied: tzlocal in c:\users\admin\anaconda3\lib\site-packages (from dateparser<2.0.0,>=1.0.0->facebook_scraper) (5.2)
Requirement already satisfied: six<=1.5 in c:\users\admin\anaconda3\lib\site-packages (from python-dateutil<=2.8.1->pandas) (1.16.0)
Requirement already satisfied: requests in c:\users\admin\anaconda3\lib\site-packages (from requests-html<0.11.0,>=0.10.0->facebook_scraper) (2.31.0)
Requirement already satisfied: pyquery in c:\users\admin\anaconda3\lib\site-packages (from requests-html<0.11.0,>=0.10.0->facebook_scraper) (2.0.0)
```

1.2 File www.facebook.cookies.txt:

Dùng extension Get cookies.txt trên chrome



Nhấn Export và lưu vào tệp data

Get cookies.txt for <https://www.facebook.com/Liverpool...>

Export Export As Copy

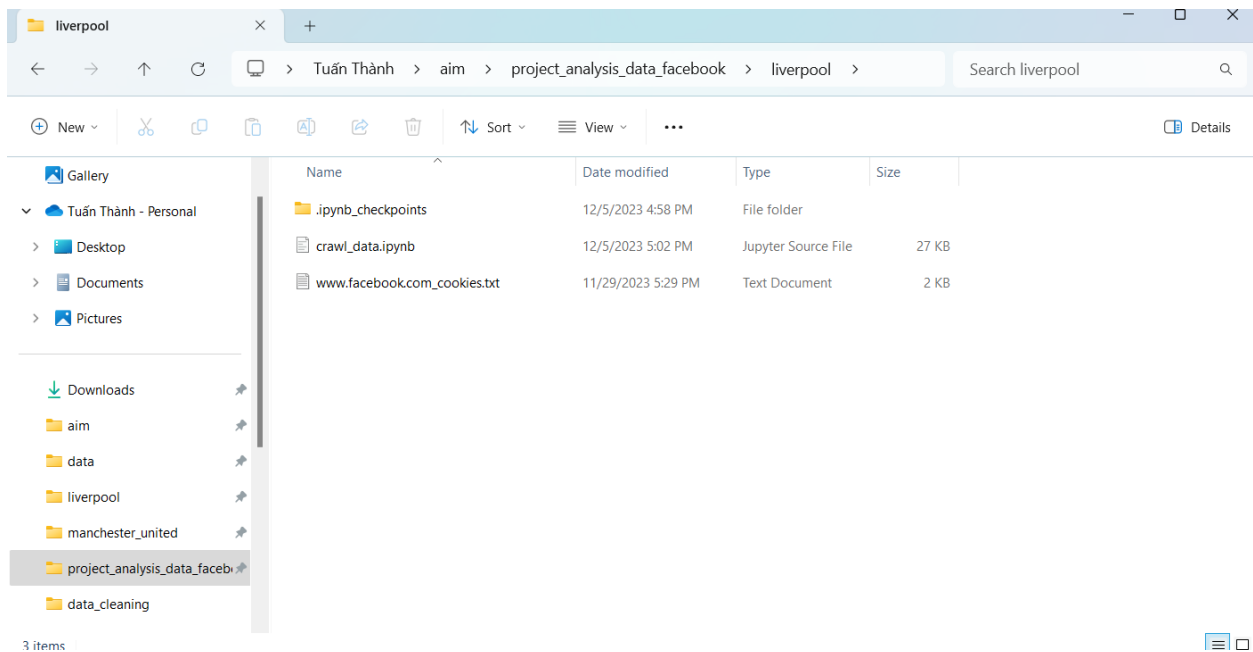
Export All Cookies

Export Format: Netscape Table Nowrap: ☒

Domain	Include Subdomains	Path	Secure	Expiry	Name	
.facebook.com	TRUE	/	TRUE	1735974341	sb	d-pRZUzll
.facebook.com	TRUE	/	TRUE	1734695109	datr	wgBWZQ:
.facebook.com	TRUE	/	TRUE	1701704320	locale	vi_VN
.www.facebook.com	TRUE	/	TRUE	1735793265	m_ls	%7B%22c
.facebook.com	TRUE	/	TRUE	1733119015	c_user	10007958
.facebook.com	TRUE	/	TRUE	1702188229	dpr	1.1000000
.facebook.com	TRUE	/	TRUE	1733119015	xs	47%3Asxr
.facebook.com	TRUE	/	TRUE	1709359426	fr	1gJalXjVf
.facebook.com	TRUE	/	TRUE	0	presence	C%7B%22
.facebook.com	TRUE	/	TRUE	1702188265	wd	1164x808

Buy Developer a Coffee

Lưu cookies vào trong thư mục có tên đội bóng.



1.3 Thu thập dữ liệu về:

Dùng thư viện facebook_scraper.

Trong link GitHub file thu thập dữ liệu được lưu trong thư mục liverpool

Khai báo thư viện

```
1 from facebook_scraper import get_posts
2 import pandas as pd
3 import numpy as np
```

Khai báo đường link, đường dẫn và số bài viết thu thập

```
1 FANPAGE_LINK = "LiverpoolFC"
2 FOLDER_PATH = "C:/Users/Admin/aim/project_analysis_data_facebook/liverpool/data/"
3 COOKIE_PATH = "C:/Users/Admin/aim/project_analysis_data_facebook/liverpool/data/www.facebook.com_c
4
5 PAGES_NUMBER = 10
```

Tiến hành thu thập dữ liệu

```
1 post_list = []
2 for post in get_posts(FANPAGE_LINK,
3                       options={"comments": True, "reactions": True, "allow_extra_requests": True},
4                       extra_info=True, pages=PAGES_NUMBER, cookies=COOKIE_PATH):
5     post_list.append(post)
6     print(post)
```

```
{'post_id': '904290937732405', 'text': '"You want to win every game and want to do your best every game. Hopefully as a by-product of that we do qualify." \n\nBen Doak on #UEL chances, development, Mo Salah guidance, and more \n\n"Bạn muốn chiến thắng mỗi trò chơi và muốn nỗ lực hết mình mỗi trò chơi. Hy vọng là một sản phẩm phụ mà chúng tôi có đủ tiêu chuẩn. " \n\nBen Doak về #UEL cơ hội, phát triển, hướng dẫn Mo Salah và nhiều hơn thế nữa \n\nLIVERPOOLFC.COM\nBen Doak on Europa League chances, development, Mo Salah guidance and more - Liverpool FC', 'post_text': '"You want to win every game and want to do your best every game. Hopefully as a by-product of that we do qualify." \n\nBen Doak on #UEL chances, development, Mo Salah guidance, and more \n\n"Bạn muốn chiến thắng mỗi trò chơi và muốn nỗ lực hết mình mỗi trò chơi. Hy vọng là một sản phẩm phụ mà chúng tôi có đủ tiêu chuẩn. " \n\nBen Doak về #UEL cơ hội, phát triển, hướng dẫn Mo Salah và nhiều hơn thế nữa \n\n', 'shared_text': 'LIVERPOOLFC.COM\nBen Doak on Europa League chances, development, Mo Salah guidance and more - Liverpool FC', 'original_text': '"You want to win every game and want to do your best every game. Hopefully as a by-product of that we do qualify." \n\nBen Doak on #UEL chances, development, Mo Salah guidance, and more \n\n', 'time': datetime.datetime(2023, 11, 29, 16, 33, 33), 'timestamp': 1701250413, 'image': None, 'image_lowquality': 'https://external.fhan5-2.fna.fbcdn.net/emg1/v/t13/17452639050244043995?url=https%3A%2F%2Fbackend.liverpoolfc.com%2Fsites%2Fdefault%2Ffiles%2Fstyles%2Fflg%2Fpublic%2F2023-11%2Fben-doak-liverpool-fc-281123.jpg%3Fitok%3DG80zaaje&fb_obo=1&utld=liverpoolfc.com&stp=c0.5000x0.5000f_dst-jpg_flffffff_p476x249_q75&ccb=13-1&oh=06_AbFguQb4X1yIrCD8Ssvo54qagyXAKlbxK4vx5KCw9civdQ&oe=6568B12C&nc_sid=2b8b93', 'images': [], 'images_description': [], 'images_lowquality': 'https://external.fhan5-2.fna.fbcdn.net/emg1/v/t13/17452639050244043995?url=https%3A%2F%2Fbackend.liverpoolfc.com%2Fsites%2Fdefault%2Ffiles%2Fstyles%2Fflg%2Fpublic%2F2023-11%2Fben-doak-liverpool-fc-281123.jpg%3Fitok%3DG80zaaje&fb_obo=1&utld=liverpoolfc.com&stp=c0.5000x0.5000f_dst-jpg_flffffff_p476x249_q75&ccb=13-1&oh=06_AbFguQb4X1yIrCD8Ssvo54qagyXAKlbxK4vx5KCw9civdQ&oe=6568B12C&nc_sid=2b8b93'}
```

1.4 Lưu liệu thu thập được:

Lưu lại kết quả thu thập từ trang page vào tệp CSV ở trong tong thư mục data

```
1 # Initialize dataframe to scrape Facebook post
2 post_df_full = pd.DataFrame(columns=post_list[0].keys(), index=range(len(post_list)), data=post_list)
3
4 # To df
5 path=FOLDER_PATH + FANPAGE_LINK + ".csv"
6 post_df_full.to_csv(path, index=False)
7 print(path)
```

C:/Users/Admin/aim/project_analysis_data_facebook/liverpool/data/LiverpoolFC.csv

1.5 Xem thông tin dữ liệu vừa thu thập được:

1	post_list
---	-----------

```
[{'post_id': '904290937732405',
  'text': '"You want to win every game and want to do your best every game. Hopefully as a by-product of that we do qualify."
  🇵🇸\n\nBen Doak on #UEL chances, development, Mo Salah guidance, and more 🇧🇷\n\n"Bạn muốn chiến thắng mỗi trò chơi và muốn nỗ lực hết mình mỗi trò chơi. Hy vọng là một sản phẩm phụ mà chúng tôi có đủ tiêu chuẩn." 🇵🇸\n\nBen Doak về #UEL cơ hội, phát triển, hướng dẫn Mo Salah và nhiều hơn thế nữa 🇧🇷\n\nLIVERPOOLFC.COM\nBen Doak on Europa League chances, development, Mo Salah guidance and more - Liverpool FC',
  'post_text': '"You want to win every game and want to do your best every game. Hopefully as a by-product of that we do qualify."
  🇵🇸\n\nBen Doak on #UEL chances, development, Mo Salah guidance, and more 🇧🇷\n\n"Bạn muốn chiến thắng mỗi trò chơi và muốn nỗ lực hết mình mỗi trò chơi. Hy vọng là một sản phẩm phụ mà chúng tôi có đủ tiêu chuẩn." 🇵🇸\n\nBen Doak về #UEL cơ hội, phát triển, hướng dẫn Mo Salah và nhiều hơn thế nữa 🇧🇷',
  'shared_text': 'LIVERPOOLFC.COM\nBen Doak on Europa League chances, development, Mo Salah guidance and more - Liverpool F
```

1	df.head()
---	-----------

	post_id	text	post_text	shared_text	original_text	time	timestamp	image	in
0	8.992690e+14	🇵🇸🇵🇸 Emile Heskey is the latest to tell the tal...	🇵🇸🇵🇸 Emile Heskey is the latest to tell the tal...	LIVERPOOLFC.COM\nMy Liverpool Story... with Emil...	🇵🇸🇵🇸 Emile Heskey is the latest to tell the tal...	11/21/2023 0:01	1700499693	NaN	https: 2.fna.fbc
1	8.992110e+14	Back-to-back PFA Premier League Fans' Player o...	Back-to-back PFA Premier League Fans' Player o...	LIVERPOOLFC.COM\nMohamed Salah named October's...	Back-to-back PFA Premier League Fans' Player o...	11/20/2023 22:00	1700492458	NaN	https: 2.fna.fbc
2	8.990810e+14	Ryan's first goal in Red 🇵🇸\n\nA look back at o...	Ryan's first goal in Red 🇵🇸\n\nA look back at o...	NaN	Ryan's first goal in Red 🇵🇸\n\nA look back at o...	11/20/2023 21:10	1700489427	NaN	https: 2.fna.fb
3	8.991520e+14	Well in, Robbo 🇵🇸 #Euro2024 🇵🇸\n\nVàng, Robbo 🇵🇸 ...	Well in, Robbo 🇵🇸 #Euro2024 🇵🇸\n\nVàng, Robbo 🇵🇸 ...	NaN	Well in, Robbo 🇵🇸 #Euro2024 🇵🇸 ...	11/20/2023 20:01	1700485284	NaN	https: 2.fna.fb
4	8.991290e+14	Two years ago today... Diogo 🇵🇸\n\nNgày này hai...	Two years ago today... Diogo 🇵🇸\n\nNgày này hai...	NaN	Two years ago today... Diogo 🇵🇸	11/20/2023 19:20	1700482820	NaN	https: 2.fna.fb

5 rows × 51 columns


```

1 df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100 entries, 0 to 99
Data columns (total 51 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   post_id                               100 non-null    float64
1   text                                  98 non-null     object
2   post_text                             98 non-null     object
3   shared_text                           21 non-null     object
4   original_text                         76 non-null     object
5   time                                  100 non-null    object
6   timestamp                             100 non-null    int64
7   image                                 37 non-null     object
8   image_lowquality                      100 non-null    object
9   images                                100 non-null    object
10  images_description                    100 non-null    object
11  images_lowquality                    100 non-null    object
12  images_lowquality_description         100 non-null    object
13  video                                 41 non-null     object
14  video_duration_seconds                0 non-null      float64
15  video_height                          0 non-null      float64
16  video_id                              41 non-null     float64
17  video_quality                         0 non-null      float64
18  video_size_MB                         0 non-null      float64
19  video_thumbnail                      41 non-null     object
20  video_watches                         0 non-null      float64
21  video_width                           0 non-null      float64
22  likes                                 0 non-null      float64
23  comments                              100 non-null    int64
24  shares                                100 non-null    int64
25  post_url                              100 non-null    object
26  link                                  23 non-null     object
27  links                                 98 non-null     object
28  user_id                              100 non-null    float64
29  username                              100 non-null    object
30  user_url                              100 non-null    object
31  is_live                              100 non-null    bool
32  factcheck                             0 non-null      float64
33  shared_post_id                        0 non-null      float64
34  shared_time                           0 non-null      float64
35  shared_user_id                        0 non-null      float64
36  shared_username                       0 non-null      float64
37  shared_post_url                       0 non-null      float64
38  available                             100 non-null    bool
39  comments_full                         100 non-null    object
40  reactors                              98 non-null     object
41  w3_fb_url                             99 non-null     object
42  reactions                             100 non-null    object
43  reaction_count                        100 non-null    int64
44  with                                  2 non-null      object
45  page_id                              100 non-null    int64
46  sharers                               0 non-null      float64
47  image_id                              23 non-null     float64
48  image_ids                             100 non-null    object
49  was_live                              100 non-null    bool
50  fetched_time                          99 non-null     object
dtypes: bool(3), float64(18), int64(5), object(25)
memory usage: 37.9+ KB

```

Phần 2 : Làm sạch và xử lý dữ liệu:

Tất cả file liên quan phần này nằm trong thư mục data_cleaning.

2.1 Chọn cột cần dùng và loại bỏ giá trị trống:

Code được cho phần này ở tệp select_column.ipynb trong thư mục data




Khai báo thư viện

```
1 import pandas as pd
```

Đọc file LiverpoolFC.csv

```
1 df = pd.read_csv('C:/Users/Admin/aim/project_analysis_data_facebook/data/LiverpoolFC.csv')
```

```
1 df
```

	post_id	text	post_text	shared_text	original_text	time	timestam
0	8.992690e+14	 Emile Heskey is the latest to tell the tal...	 Emile Heskey is the latest to tell the tal...	LIVERPOOLFC.COM\nMy Liverpool Story... with Emil...	 Emile Heskey is the latest to tell the tal...	11/21/2023 0:01	170049969
1	8.992110e+14	Back-to-back PFA Premier League Fans' Player o...	Back-to-back PFA Premier League Fans' Player o...	LIVERPOOLFC.COM\nMohamed Salah named October's...	Back-to-back PFA Premier League Fans' Player o...	11/20/2023 22:00	170049245
2	8.990810e+14	Ryan's first goal in Red 🍒\n\nA look back at o...	Ryan's first goal in Red 🍒\n\nA look back at o...	NaN	Ryan's first goal in Red 🍒\n\nA look back at o...	11/20/2023 21:10	170048942

Chọn cột

```
1 new_df = df[["post_id", "post_text", "time", "comments", "comments_full", "shares", "reactions", "r
```

```
new_df = df[["post_id", "post_text", "time", "comments", "comments_full", "shares", "reactions", "reaction_count"]]
```

```
1 new_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100 entries, 0 to 99
Data columns (total 8 columns):
#   Column          Non-Null Count  Dtype
---  -
0   post_id         100 non-null   float64
1   post_text       98 non-null    object
2   time            100 non-null   object
3   comments        100 non-null   int64
4   comments_full   100 non-null   object
5   shares          100 non-null   int64
6   reactions       100 non-null   object
7   reaction_count  100 non-null   int64
dtypes: float64(1), int64(3), object(4)
memory usage: 6.4+ KB
```

Bỏ hàng có giá trị rỗng

```
1 new_df_01 = new_df.dropna()
```

```
1 new_df_01.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 98 entries, 0 to 99
Data columns (total 8 columns):
#   Column          Non-Null Count  Dtype  
---  -
0   post_id         98 non-null    float64
1   post_text       98 non-null    object  
2   time            98 non-null    object  
3   comments        98 non-null    int64   
4   comments_full   98 non-null    object  
5   shares          98 non-null    int64   
6   reactions       98 non-null    object  
7   reaction_count  98 non-null    int64   
dtypes: float64(1), int64(3), object(4)
memory usage: 6.9+ KB
```

Lưu lại vào tệp Lp_select_column.csv

```
new_df_01.to_csv('C:/Users/Admin/aim/project_analysis_data_facebook/data/Lp_select_column.csv',
index = False)
```

2.2 Tách cột biểu cảm thành các trường cần phân tích:


code này được lưu trong thư mục data_cleaning tệp tên là divide_reactions.ipynb

Đọc file Lp_select_column.csv

```
1 import pandas as pd
2 import ast
```

```
1 df = pd.read_csv('C:/Users/Admin/aim/project_analysis_data_facebook/data/Lp_select_column.csv')
```

```
1 df
```

	post_id	post_text	time	comments	comments_full	shares	reactions	reaction_count
0	8.992690e+14	 Emile Heskey is the latest to tell the tal...	11/21/2023 0:01	16	{'comment_id': '1186163892166613', 'comment_u...	101	{'thích': 982, 'yêu thích': 183, 'thương thươn...	1175
1	8.992110e+14	Back-to-back PFA Premier League Fans' Player o...	11/20/2023 22:00	83	{'comment_id': '277210241476921', 'comment_ur...	31	{'thích': 3638, 'yêu thích': 902, 'haha': 1}	4594

Tạo 1 list chứa các dictionary của các loại reaction của từng post và lưu vào 1 list

```
1 react_list = []
2
3 for react_str in df.reactions:
4     react_dict = ast.literal_eval(react_str)
5     react_list.append(react_dict)
```

Tạo dataframe từ react_list

```
1 df_reactions = pd.DataFrame(react_list)
2 df_reactions
```

	thích	yêu thích	thương thương	haha	wow	buồn	phấn nộ
0	982	183	10	NaN	NaN	NaN	NaN
1	3638	902	49	1.0	4.0	NaN	NaN
2	4072	1022	55	3.0	4.0	1.0	NaN
3	12735	3112	146	19.0	6.0	1.0	NaN
4	11699	2506	125	45.0	26.0	1.0	1.0
...
93	11518	1523	77	9.0	7.0	2.0	3.0
94	7779	934	85	3.0	9.0	39.0	4.0
95	13364	2481	174	8.0	13.0	7.0	3.0
96	26432	7266	385	15.0	6.0	2.0	1.0
97	39460	6610	353	12.0	38.0	1.0	2.0

98 rows × 7 columns

Thay thế các giá trị NaN bằng 0

```
1 df_reactions = df_reactions.fillna(0)
```


Bỏ thuộc tính reactions cũ

```
1 no_react_df = df.drop(['reactions'], axis = 1)
```

Thay thuộc tính reactions bằng dataframe df_reactions

```
1 new_df = pd.concat([no_react_df, df_reactions], axis = 1)
```

```
1 new_df
```

	post_id	post_text	time	comments	comments_full	shares	reaction_count	thích	yêu thích	thương
0	8.992690e+14	 Emile Heskey is the latest to tell the tal...	11/21/2023 0:01	16	[{'comment_id': '1186163892166613', 'comment_u...	101	1175	982	183	
1	8.992110e+14	Back-to-back PFA Premier League Fans' Player o...	11/20/2023 22:00	83	[{'comment_id': '277210241476921', 'comment_ur...	31	4594	3638	902	

Lưu lại vào file Lp_divide_reactions.csv

```
1 new_df.to_csv('C:/Users/Admin/aim/project_analysis_data_facebook/data/Lp_select_column.csv', index
```


2.3 Tách cột thời gian thành cột ngày tháng năm , thứ , giờ:

Code của phần này nằm trong tệp divide_time.ipynb của thư mục data_cleaning

Mở file Lp_divide_reactions.csv

```
1 import pandas as pd
2 from datetime import datetime

1 df = pd.read_csv('C:/Users/Admin/aim/project_analysis_data_facebook/data/Lp_select_column.csv')
```

	post_id	post_text	time	comments	comments_full	shares	reaction_count	thích	yêu thích	thương
0	8.992690e+14	 Emile Heskey is the latest to tell the tal...	11/21/2023 0:01	16	[{'comment_id': '1186163892166613', 'comment_ur...	101	1175	982	183	
1	8.992110e+14	Back-to-back PFA Premier League Fans' Player o...	11/20/2023 22:00	83	[{'comment_id': '277210241476921', 'comment_ur...	31	4594	3638	902	
2	8.990810e+14	Ryan's first goal in Red 🍅\n\nA look back at o...	11/20/2023 21:10	49	[{'comment_id': '388506910181825', 'comment_ur...	53	5157	4072	1022	
3	8.991520e+14	Well in, Robbo 🍌 #Euro2024 \n\nVâng, Robbo 🍌 ...	11/20/2023 20:01	150	[{'comment_id': '314996431472511', 'comment_ur...	61	16019	12735	3112	1
4	8.991290e+14	Two years ago today... Diogo 🍌\n\nNgày này hai...	11/20/2023 19:20	211	[{'comment_id': '1028198778438522', 'comment_u...	211	14403	11699	2506	1

Tách time ra ngày tháng năm

Ép kiểu thuộc tính time từ str thành datetime

```
1 date_list = []
2
3 for date in df.time:
4     date_list.append(date[:10])
```

tạo 1 list chứa các giờ đăng bài

```
1 hour_list = []
2
3 for time in df.time:
4     time_obj = datetime.strptime(time, '%m/%d/%Y %H:%M')
5     time = time_obj.time()
6     hour_list.append(time.hour)
```

```
1 hour_list
```

[0, 22, 21, 20, 19, 18, 17, 16, 15, 5, 3, 1, 0, 23, 22, 21, 20, 18, 17, 15, 14, 13, 12, 11, 10, 9, 8, 7, 6, 4, 2]

Tạo dataframe chứa ngày và dataframe chứa giờ đăng bài

```
1 df_date = pd.DataFrame(date_list, columns = ["date"])
2 df_hour = pd.DataFrame(hour_list, columns = ["hour"])
3 df_hour_date = pd.concat([df_hour, df_date], axis = 1)
4 # bỏ cột time trong dataframe cũ
5 no_time_df = df.drop(['time'], axis = 1)
6 # nối dataframe đã xóa time với df_hour_date
7 new_df = pd.concat([no_time_df, df_hour_date], axis = 1)
```

```
1 new_df
```

1	new_df													
t_text	comments	comments_full	shares	reaction_count	thích	yêu thích	thương thương	haha	wow	buồn	phấn nộ	hour	date	
eskey ell the tal...	16	[{'comment_id': '1186163892166613', 'comment_u...	101	1175	982	183	10	0.0	0.0	0.0	0.0	0	11/21/2023	
k PFA Fans' er o...	83	[{'comment_id': '277210241476921', 'comment_ur...	31	4594	3638	902	49	1.0	4.0	0.0	0.0	22	11/20/2023	
goal in A look at o...	49	[{'comment_id': '388506910181825', 'comment_ur...	53	5157	4072	1022	55	3.0	4.0	1.0	0.0	21	11/20/2023	
bo 🍌 o2024 Robbo 🍌 ...	150	[{'comment_id': '314996431472511', 'comment_ur...	61	16019	12735	3112	146	19.0	6.0	1.0	0.0	20	11/20/2023	
oday... nNgày / hai...	211	[{'comment_id': '1028198778438522', 'comment_u...	211	14403	11699	2506	125	45.0	26.0	1.0	1.0	19	11/20/2023	
...	
nk him period he ...	97	[{'comment_id': '1120275918934925', 'comment_u...	43	13139	11518	1523	77	9.0	7.0	2.0	3.0	17	11/13/2023	
Klopp omez ima ...	114	[{'comment_id': '707770537958734', 'comment_ur...	34	8853	7779	934	85	3.0	9.0	39.0	4.0	16	11/13/2023	

Tạo 1 list chứa thứ đăng bài

1	day_of_week_list = []
2	
3	for date in new_df.date:
4	specific_date = pd.to_datetime(date)
5	day_of_week = specific_date.strftime('%A')
6	day_of_week_list.append(day_of_week)

Thêm thứ vào new_df

1	day_df = pd.DataFrame(day_of_week_list, columns = ["day"])
2	new_df_2 = pd.concat([new_df, day_df], axis = 1)

1	day_df
---	--------

day	
0	Tuesday
1	Monday
2	Monday
3	Monday
4	Monday
...	...
93	Monday
94	Monday
95	Monday
96	Monday
97	Tuesday

98 rows × 1 columns

1	new_df_2
---	----------

mments	comments_full	shares	reaction_count	thích	yêu thích	thương thương	haha	wow	buồn	phấn nộ	hour	date	day
16	['comment_id': '1186163892166613', 'comment_ur...	101	1175	982	183	10	0.0	0.0	0.0	0.0	0	11/21/2023	Tuesday
83	['comment_id': '277210241476921', 'comment_ur...	31	4594	3638	902	49	1.0	4.0	0.0	0.0	22	11/20/2023	Monday
49	['comment_id': '388506910181825', 'comment_ur...	53	5157	4072	1022	55	3.0	4.0	1.0	0.0	21	11/20/2023	Monday
150	['comment_id': '314996431472511', 'comment_ur...	61	16019	12735	3112	146	19.0	6.0	1.0	0.0	20	11/20/2023	Monday
211	['comment_id': '1028198778438522', 'comment_ur...	211	14403	11699	2506	125	45.0	26.0	1.0	1.0	19	11/20/2023	Monday
...
97	['comment_id': '1120275918934925', 'comment_ur...	43	13139	11518	1523	77	9.0	7.0	2.0	3.0	17	11/13/2023	Monday
114	['comment_id': '707770537958734', 'comment_ur...	34	8853	7779	934	85	3.0	9.0	39.0	4.0	16	11/13/2023	Monday

Lưu vào file data_ready.csv

1	new_df_2.to_csv('C:/Users/Admin/aim/project_analysis_data_facebook/data/data_ready.csv', index=False)
---	---

Phần 3 : Phân tích dữ liệu:

Tất cả code phần này nằm trong thư mục analysis.

3.1 Phân tích nội dung:








Code phần này nằm trong tệp analysis_content.ipynb

Đọc file data_ready.csv

```
1 import pandas as pd
2 import matplotlib.pyplot as plt
```

```
1 df = pd.read_csv('C:/Users/Admin/aim/project_analysis_data_facebook/data/data_ready.csv')
```

```
1 df.head()
```

	post_id	post_text	comments	comments_full	shares	reaction_count	thích	yêu thích	thương thương	haha	wow	buồn
0	8.992690e+14	  Emile Heskey is the latest to tell the tal...	16	[{'comment_id': '1186163892166613', 'comment_u...	101	1175	982	183	10	0.0	0.0	0.0
1	8.992110e+14	Back-to-back PFA Premier League Fans' Player O...	83	[{'comment_id': '277210241476921', 'comment_ur...	31	4594	3638	902	49	1.0	4.0	0.0
2	8.990810e+14	Ryan's first goal in Red  \n\nA look back at O...	49	[{'comment_id': '388506910181825', 'comment_ur...	53	5157	4072	1022	55	3.0	4.0	1.0
3	8.991520e+14	Well in, Robbo  #Euro2024  \n\nVàng, Robbo  ...	150	[{'comment_id': '314996431472511', 'comment_ur...	61	16019	12735	3112	146	19.0	6.0	1.0
4	8.991290e+14	Two years ago today... Diogo  \n\nNgày này hai...	211	[{'comment_id': '1028198778438522', 'comment_u...	211	14403	11699	2506	125	45.0	26.0	1.0

Lấy toàn bộ nội dung các bài đăng

```
: 1 list_text_full = str(df['post_text'])
```

Word cloud

```
1 from wordcloud import WordCloud, STOPWORDS
2
3 wordcloud = WordCloud(stopwords=STOPWORDS,
4                       background_color='white',
5                       max_words=300,
6                       width=2000, height=1200
7                       ).generate(list_text_full)
8 plt.figure(figsize=(40,20))
9 plt.clf()
10 plt.imshow(wordcloud)
11 plt.axis('off')
12 plt.show()
```



Nhìn biểu đồ trên ta thấy từ back xuất hiện nhiều nhất.

Các nội dung bài đăng hầu như xoay quanh huấn luyện viên, tên các cầu thủ và sân vận động của Liverpool FC

3.2 Phân tích bình luận:















Code phần này nằm trong tệp analysis_comments.ipynb

Đọc file data_ready.csv

```
1 import pandas as pd
2 import matplotlib.pyplot as plt
3 import numpy as np
```

```
1 df = pd.read_csv('C:/Users/Admin/aim/project_analysis_data_facebook/data/data_ready.csv')
```

```
1 df
```

	post_id	post_text	comments	comments_full	shares	reaction_count	thích	yêu thích	thương thương	haha	wow	buồn	phẫn nộ	hour	date
0	8.992690e+14	 Emile Heskey is the latest to tell the tal...	16	[{"comment_id": "1186163892166613", "comment_u...	101	1175	982	183	10	0.0	0.0	0.0	0.0	0	11/21/202
1	8.992110e+14	Back-to-back PFA Premier League Fans' Player o...	83	[{"comment_id": "277210241476921", "comment_ur...	31	4594	3638	902	49	1.0	4.0	0.0	0.0	22	11/20/202
2	8.990810e+14	Ryan's first goal in Red  A look back at o...	49	[{"comment_id": "388506910181825", "comment_ur...	53	5157	4072	1022	55	3.0	4.0	1.0	0.0	21	11/20/202
3	8.991520e+14	Well in, Robbo  #Euro2024  Vàng, Robbo  ...	150	[{"comment_id": "314996431472511", "comment_ur...	61	16019	12735	3112	146	19.0	6.0	1.0	0.0	20	11/20/202
4	8.991290e+14	Two years ago today... Diogo   Ngây này hai...	211	[{"comment_id": "1028198778438522", "comment_u...	211	14403	11699	2506	125	45.0	26.0	1.0	1.0	19	11/20/202
...
93	8.948540e+14	"I really thank him because in this period he ...	97	[{"comment_id": "1120275918934925", "comment_u...	43	13139	11518	1523	77	9.0	7.0	2.0	3.0	17	11/13/202
94	8.948430e+14	Jürgen Klopp confirmed Joe Gomez and Ibrahima ...	114	[{"comment_id": "707770537958734", "comment_ur...	34	8853	7779	934	85	3.0	9.0	39.0	4.0	16	11/13/202
95	8.948350e+14	Anfield remembers  Anfield nhớ.	149	[{"comment_id": "7303758222967701", "comment_u...	309	16050	13364	2481	174	8.0	13.0	7.0	3.0	16	11/13/202
96	8.944540e+14	Good morning    Chào buổi sáng  	417	[{"comment_id": "1483340398900267", "comment_u...	124	34107	26432	7266	385	15.0	6.0	2.0	1.0	15	11/13/202
97	8.955110e+14	Big Virg against Brentford   Big Virg chống...	631	[{"comment_id": "304499032539488", "comment_ur...	727	46476	39460	6610	353	12.0	38.0	1.0	2.0	18	11/14/202

98 rows x 16 columns

Thêm cột Index

```
1 df = df.reset_index(inplace = False)
2 df["index"] = df["index"] + 1
```

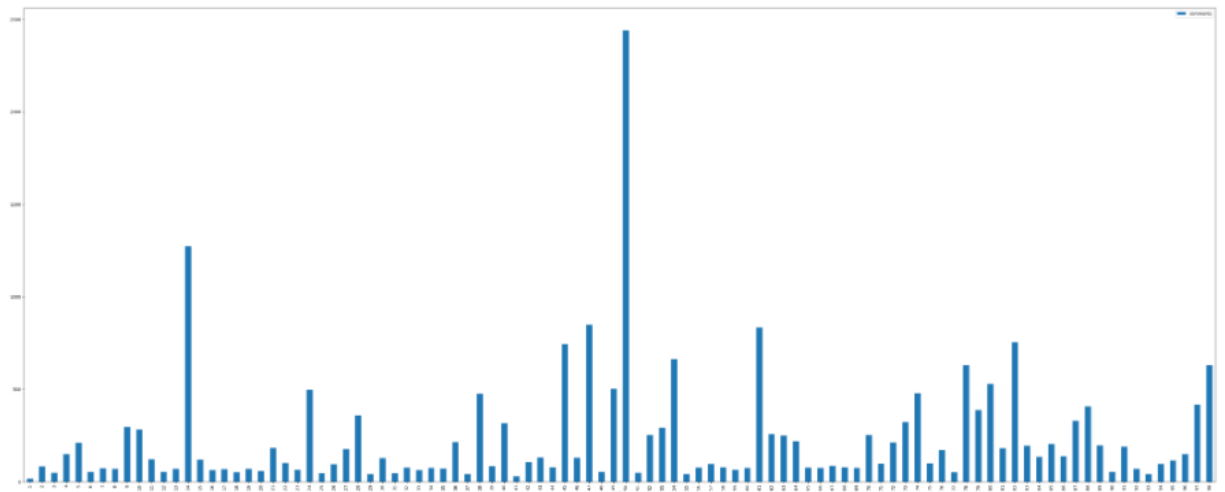
```
1 df
```

	index	post_id	post_text	comments	comments_full	shares	reaction_count	thích	yêu thích	thương thương	haha	wow	buồn	phấn nộ	hour
0	1	8.992690e+14	🔴👤 Emile Heskey is the latest to tell the tal...	16	'1186163892166613', 'comment_u...	101	1175	982	183	10	0.0	0.0	0.0	0.0	0
1	2	8.992110e+14	Back-to-back PFA Premier League Fans' Player o...	83	'277210241476921', 'comment_ur...	31	4594	3638	902	49	1.0	4.0	0.0	0.0	22
2	3	8.990810e+14	Ryan's first goal in Red 🔴👤\n\nA look back at o...	49	'388506910181825', 'comment_ur...	53	5157	4072	1022	55	3.0	4.0	1.0	0.0	21
3	4	8.991520e+14	Well in, Robbo 🏆\n\n#Euro2024 🇵🇹\n\nVâng, Robbo 🏆 ...	150	'314996431472511', 'comment_ur...	61	16019	12735	3112	146	19.0	6.0	1.0	0.0	20
4	5	8.991290e+14	Two years ago today... Diogo 🏆\n\nNgày này hải...	211	'1028198778438522', 'comment_u...	211	14403	11699	2506	125	45.0	26.0	1.0	1.0	19
...
93	94	8.948540e+14	"I really thank him because in this period he ...	97	'1120275918934925', 'comment_u...	43	13139	11518	1523	77	9.0	7.0	2.0	3.0	17
94	95	8.948430e+14	Jürgen Klopp confirmed Joe Gomez and Ibrahima ...	114	'707770537958734', 'comment_ur...	34	8853	7779	934	85	3.0	9.0	39.0	4.0	16
95	96	8.948350e+14	Anfield remembers.\n\nAnfield nhớ.	149	'7303758222967701', 'comment_u...	309	16050	13364	2481	174	8.0	13.0	7.0	3.0	16
96	97	8.944540e+14	Good morning 🌞\n\nChào buổi sáng 🌞🔴	417	'1483340398900267', 'comment_u...	124	34107	26432	7266	385	15.0	6.0	2.0	1.0	15
97	98	8.955110e+14	Big Virg against Brentford 🏆\n\nBig Virg chống...	631	'304499032539488', 'comment_ur...	727	46476	39460	6610	353	12.0	38.0	1.0	2.0	18

98 rows × 17 columns

Biểu đồ số lượng comment của từng bài đăng

```
1 df.plot(kind = "bar", x = "index", y = "comments", figsize = (50, 20));
```



Kết luận: Bài viết thứ 50 là bài viết có số lượt commen nhiều nhất.

Tổng số lượng comments

```
1 sum(df.comments)
```

22625

Số lượng comment trung bình

```
1 df.comments.mean()
```

230.8673469387755

Số lượng comment của các bài đăng dao động trong khoảng

```
1 df.comments.median()
```

116.5

Bài viết có số lượt comment lớn nhất

```
1 df.index[df['comments'] == max(df.comments)]
```

Index([49], dtype='int64')

```
1 df.iloc[72]
```

```
index                                73
post_id                            89554800000000.0
post_text      Every angle of Diogo Jota18's fine finish from...
comments                                322
comments_full  [{'comment_id': '308242012017211', 'comment_ur...
shares                                453
reaction_count                       66650
thích                                57637
yêu thích                             8452
thương thương                         449
haha                                  17.0
wow                                  81.0
buồn                                  6.0
phấn nộ                              8.0
hour                                  1
date                                11/15/2023
day                                Wednesday
Name: 72, dtype: object
```

Bài viết có số comments nhỏ nhất

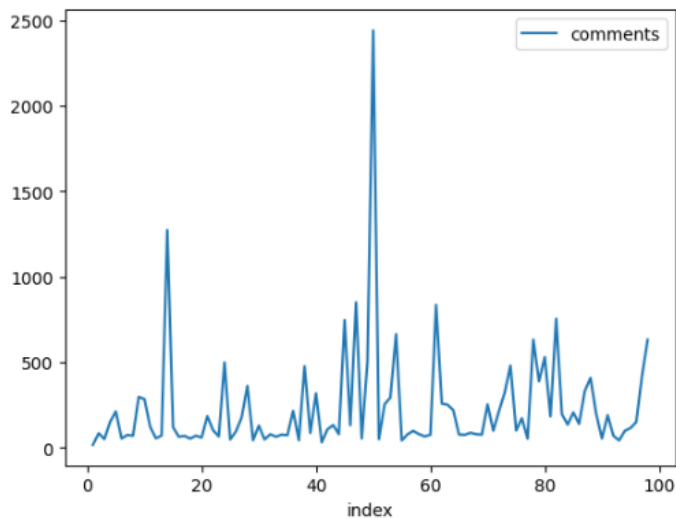
```
1 df.index[df['comments'] == min(df.comments)]
```

Index([0], dtype='int64')

Biểu đồ đường thể hiện sự biến động của trường comments:

Biến động số lượt comment qua từng bài đăng

```
1 df.plot(kind = "line", x = "index", y = "comments");
```



Qua biểu đồ ta thấy số comment của các bài không đồng đều trong tập dữ liệu.

3.3 Phân tích biểu cảm:

Code phần này nằm trong tệp analysis_reactions.ipynb

Đọc file data_ready.csv

```
1 import pandas as pd
2 import matplotlib.pyplot as plt
3 import numpy as np
```

```
1 df = pd.read_csv('C:/Users/Admin/aim/project_analysis_data_facebook/data/data_ready.csv')
```

```
1 df.info()
```

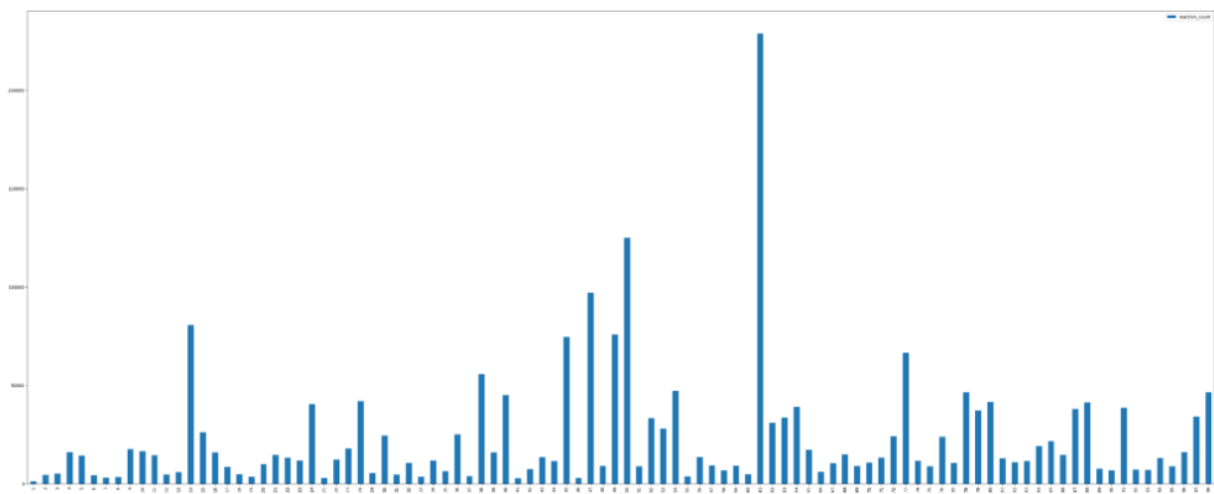
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 98 entries, 0 to 97
Data columns (total 16 columns):
 #   Column                Non-Null Count  Dtype  
---  --
 0   post_id               98 non-null    float64
 1   post_text             98 non-null    object  
 2   comments              98 non-null    int64   
 3   comments_full         98 non-null    object  
 4   shares                98 non-null    int64   
 5   reaction_count        98 non-null    int64   
 6   thích                98 non-null    int64   
 7   yêu thích            98 non-null    int64   
 8   thương thương        98 non-null    int64   
 9   haha                 98 non-null    float64  
10   wow                  98 non-null    float64  
11   buồn                 98 non-null    float64  
12   phẫn nộ              98 non-null    float64  
13   hour                 98 non-null    int64   
14   date                 98 non-null    object  
15   day                  98 non-null    object  
dtypes: float64(5), int64(7), object(4)
memory usage: 12.4+ KB
```

Thêm cột Index

```
1 df = df.reset_index(inplace = False)
2 df["index"] = df["index"] + 1
```

Biểu đồ số lượng reaction cho từng bài đăng

```
1 df.plot(kind = "bar", x = "index", y = "reaction_count", figsize = (50, 20));
```

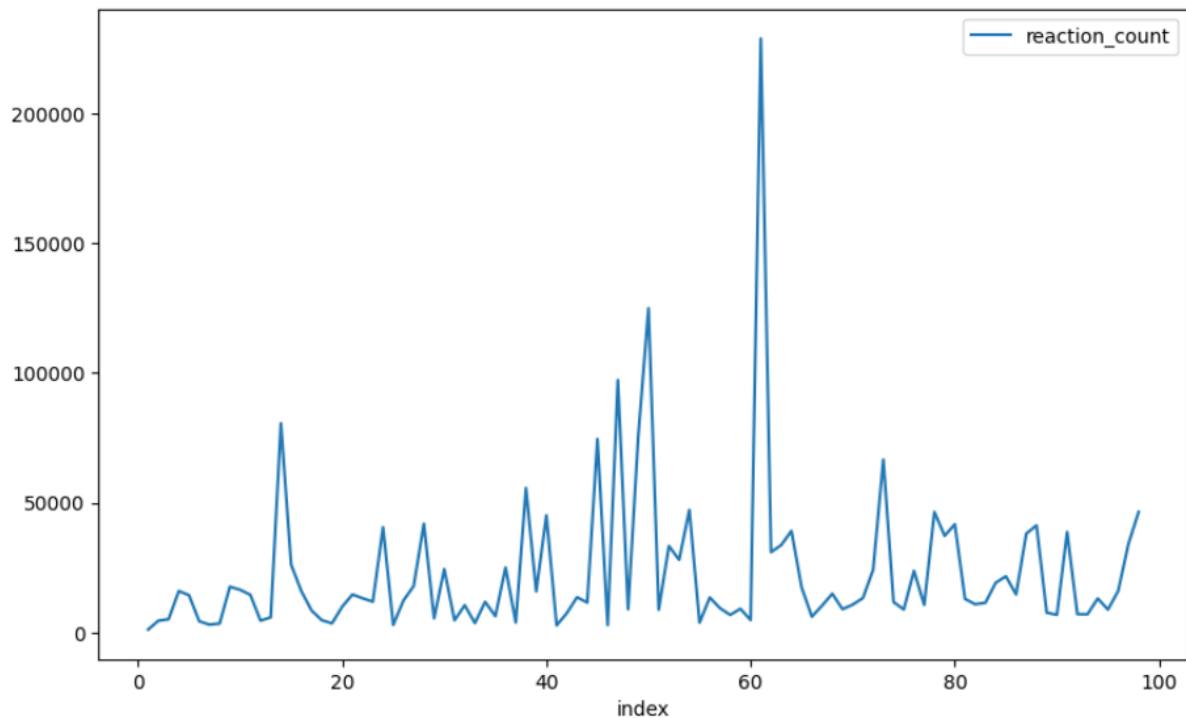


Qua biểu đồ ta thấy bài viết thứ 61 là bài viết có số lượt reaction nhiều nhất

Bài viết có số reations ít nhất là bài viết đầu tiên trong tập dữ liệu

Biến động reaction qua từng bài đăng

```
1 df.plot(kind = "line", x = "index", y = "reaction_count", figsize=(10,6));
```



Qua biểu đồ đường trên ta thấy số reactions của fanpage không đồng đều

Tổng số lượng reaction

```
1 sum(df.reaction_count)
```

2257377

Lượng reaction trung bình

```
1 df.reaction_count.mean()
```

23034.45918367347

Biến động reaction

```
1 df.reaction_count.median()
```

13180.5

Bài đăng có nhiều reaction nhất

```
1 df.index[df['reaction_count'] == max(df.reaction_count)]
```

Index([60], dtype='int64')

```
1 df.iloc[60]
```

```
index                                61
post_id                        89630500000000.0
post_text    Goal involvements in 1 5 consecutive Premie...
comments                                834
comments_full    [{'comment_id': '883159683161529', 'comment_ur...
shares                                1835
reaction_count                228877
thích                                192331
yêu thích                    33615
thương thương                    2545
haha                                102.0
wow                                228.0
buồn                                22.0
phấn nộ                    34.0
hour                                1
date                        11/16/2023
day                        Thursday
Name: 60, dtype: object
```

Bài đăng có ít reaction nhất

```
1 df.index[df['reaction_count'] == min(df.reaction_count)]
```

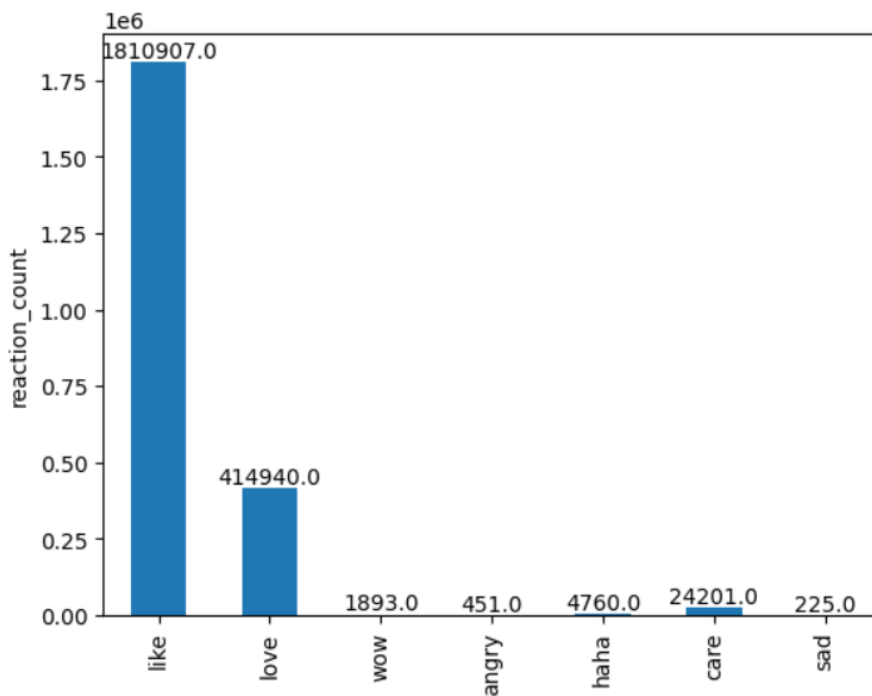
Index([0], dtype='int64')

```
1 df.iloc[0]
```

```
index                                1
post_id                        899269000000000.0
post_text    🚫 Emile Heskey is the latest to tell the tal...
comments                                16
comments_full    [{'comment_id': '1186163892166613', 'comment_u...
shares                                101
reaction_count                1175
thích                                982
yêu thích                    183
thương thương                    10
haha                                0.0
wow                                0.0
buồn                                0.0
phấn nộ                    0.0
hour                                0
date                        11/21/2023
day                        Tuesday
Name: 0, dtype: object
```

Biểu đồ lượt tương tác từng reaction

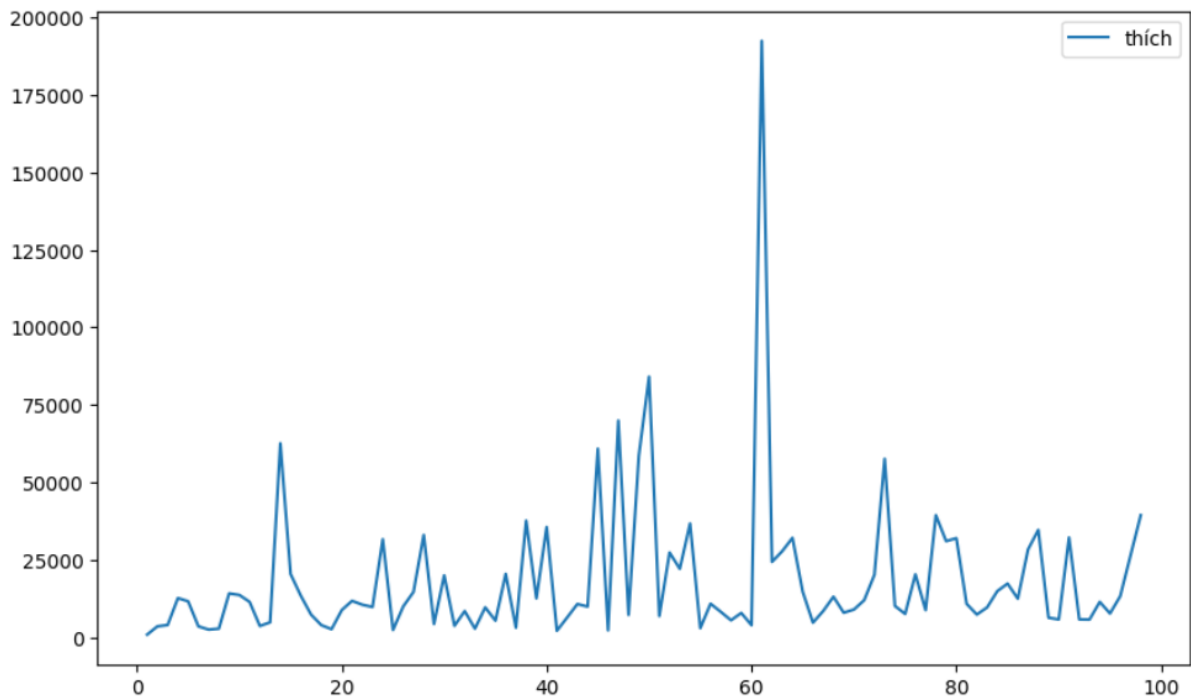
```
1 like = sum(df.thích)
2 love = sum(df['yêu thích'])
3 wow = sum(df.wow)
4 angry = sum(df['phẫn nộ'])
5 haha = sum(df.haha)
6 care = sum(df['thương thương'])
7 sad = sum(df.buồn)
8
9 react_dict = {"like": like,
10               "love": love,
11               "wow": wow,
12               "angry": angry,
13               "haha": haha,
14               "care": care,
15               "sad": sad}
16
17 react_series = pd.Series(react_dict)
18 react_plt = react_series.plot(kind = "bar", ylabel = 'reaction_count')
19
20 for b in react_plt.patches:
21     react_plt.annotate(b.get_height(), (b.get_x() + b.get_width() / 2, b.get_height()), ha = 'cent
22
23 plt.show()
```



Qua biểu đồ ta thấy đa số người xem chọn like và số lượt like vượt trội hẳn so với các trường còn lại số lượt thích.

Sự biến động của like

```
1 df.plot(kind = "line", x = "index", y = "thích", figsize=(10,6));
```



Bài đăng có số lượt thích nhiều nhất

```
1 df.index[df['thích'] == max(df.thích)]
```

Index([60], dtype='int64')

```
1 df.iloc[60]
```

```
index                                61
post_id                        89630500000000.0
post_text    Goal involvements in 1 5 consecutive Premie...
comments                                834
comments_full    [{'comment_id': '883159683161529', 'comment_ur...
shares                                1835
reaction_count                228877
thích                        192331
yêu thích                        33615
thương thương                    2545
haha                            102.0
wow                             228.0
buồn                             22.0
phấn nộ                          34.0
hour                             1
date                        11/16/2023
day                        Thursday
Name: 60, dtype: object
```

Bài đăng có số lượt thích ít nhất

```
1 df.index[df['thích'] == min(df.thích)]
```

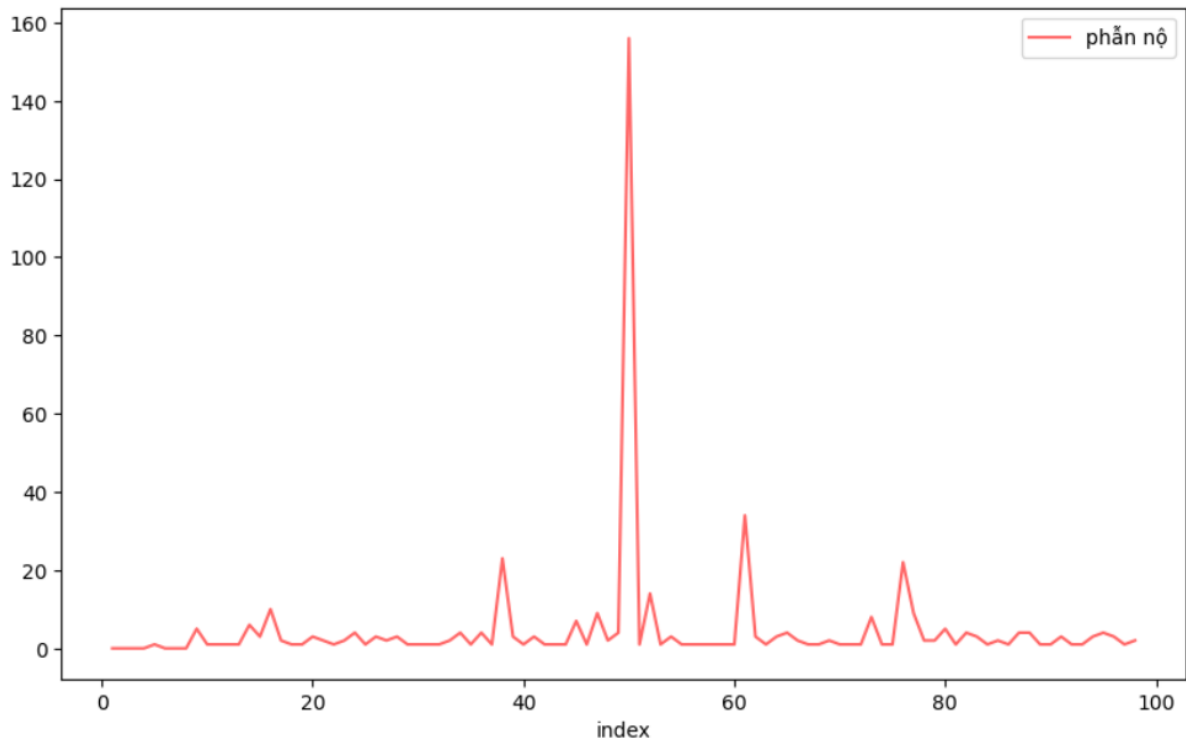
Index([0], dtype='int64')

```
1 df.iloc[0]
```

```
index                                1
post_id                        899269000000000.0
post_text      📍 Emile Heskey is the latest to tell the tal...
comments                                16
comments_full  [{'comment_id': '1186163892166613', 'comment_u...
shares                                101
reaction_count                        1175
thích                                982
yêu thích                             183
thương thương                         10
haha                                  0.0
wow                                  0.0
buồn                                  0.0
phấn nộ                              0.0
hour                                  0
date                                11/21/2023
day                                Tuesday
Name: 0, dtype: object
```

Sự biến động của phấn nộ

```
1 df.plot(kind = "line", x = "index", y = "phấn nộ", color = "#ff6666", figsize=(10,6));
```



Bài viết có nhiều phần nộ nhất

```
1 df.index[df['phần nộ'] == max(df['phần nộ'])]
```

```
Index([49], dtype='int64')
```

```
1 df.iloc[49]
```

```
index                                50
post_id                            896926000000000.0
post_text    ⚽⚽⚽⚽\n\nFour goals for Mo in Egypt's 6-0 win o...
comments                                2440
comments_full    [{'comment_id': '180509865130311', 'comment_ur...
shares                                889
reaction_count                            124991
thích                                84103
yêu thích                            37153
thương thương                            2901
haha                                493.0
wow                                175.0
buồn                                10.0
phần nộ                            156.0
hour                                1
date                                11/17/2023
day                                Friday
Name: 49, dtype: object
```

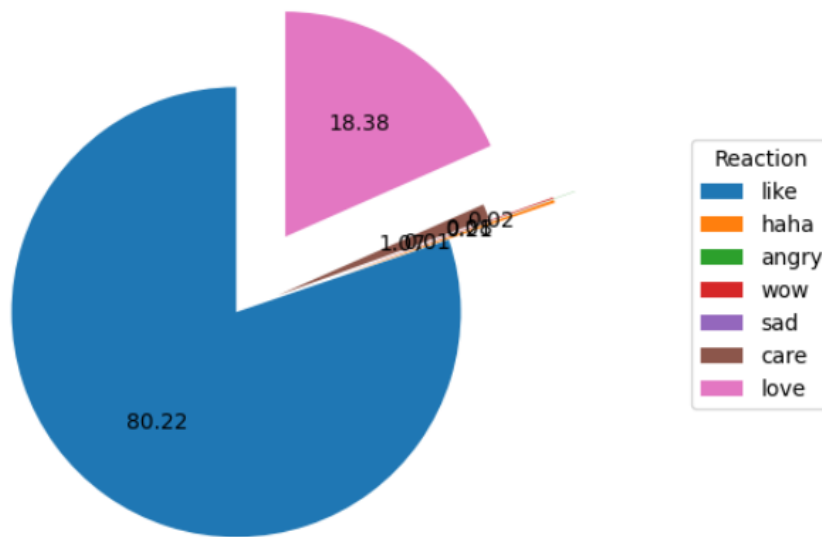
Qua biểu đồ phân tích sự biến động về số phần nộ và số phần nộ lớn nhất ta thấy khi Liverpool thắng Manchester United thì fan MU ngay lập tức vào fanpage của Liverpool thả lượt phần nộ cho bài đăng ăn mừng chiến thắng của Liverpool.

MU quá tệ!!!!!!!!!!!!

Tỉ trọng giữa các loại reaction

```
1 react_dict = {
2     'like': sum(df['thích']),
3     'haha': sum(df['haha']),
4     'angry': sum(df['phần nộ']),
5     'wow': sum(df['wow']),
6     'sad': sum(df['buồn']),
7     'care': sum(df['thương thương']),
8     'love': sum(df['yêu thích'])
9 }
10 react = []
11 number = []
12
13 explode = (0.0, 0.5, 0.6, 0.5, 0.3, 0.2, 0.4)
14
15 for x, y in react_dict.items():
16     react.append(x)
17     number.append(y)
18
19 plt.pie(number, labels=['']*len(react), autopct='%.2f', explode = explode, startangle = 90)
20
21 plt.axis('equal')
22 plt.title("Tỉ trọng các loại reaction")
23
24 plt.legend(react, title="Reaction", loc="center left", bbox_to_anchor=(1, 0.5))
25
26 plt.show()
```

Tỉ trọng các loại reaction



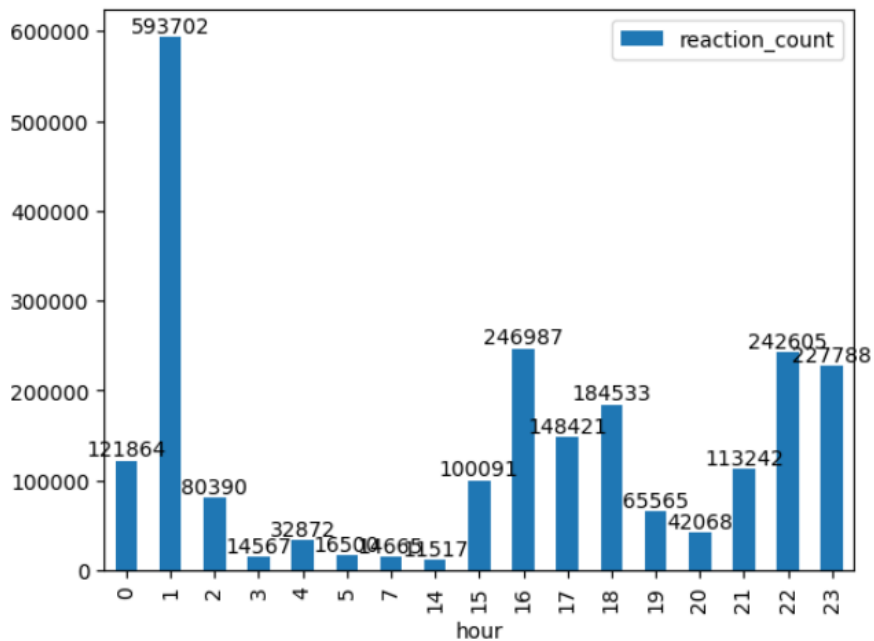
Qua biểu đồ trên ta thấy số biểu cảm haha, wow, buồn, phẫn nộ, thương thương chiếm rất ít trong tỉ trọng số biểu cảm của tất cả bài đăng.

Số reaction theo giờ đăng

```

1 hour_react_plt = df[['hour', 'reaction_count']].groupby(['hour']).sum('reaction_count').plot(kind
2 for b in hour_react_plt.patches:
3     hour_react_plt.annotate(str(b.get_height()), (b.get_x() + b.get_width() / 2., b.get_height()),
4 plt.show()

```



Do đội bóng nằm ở nước Anh nên đa số người theo dõi nằm ở Châu Âu, vì thế cái khung 1h theo giờ Việt Nam rất được nhiều người quan tâm nên 1h là giờ mà số lượt reaction nhiều nhất.

3.4 Phân tích số lượt chia sẻ:

Code phần này nằm trong thư mục analysis_shares.ipynb

Đọc file data_ready.csv

```
1 import pandas as pd
2 import matplotlib.pyplot as plt
3 import numpy as np
```

```
1 df = pd.read_csv('C:/Users/Admin/aim/project_analysis_data_facebook/data/data_ready.csv')
```

1	df										
	post_id	post_text	comments	comments_full	shares	reaction_count	thích	yêu thích	thương	haha	w
0	8.992690e+14	🔴👤 Emile Heskey is the latest to tell the tal...	16	[{'comment_id': '1186163892166613', 'comment_ur...	101	1175	982	183	10	0.0	
1	8.992110e+14	Back-to-back PFA Premier League Fans' Player o...	83	[{'comment_id': '277210241476921', 'comment_ur...	31	4594	3638	902	49	1.0	
2	8.990810e+14	Ryan's first goal in Red 🇵🇹\n\nA look back at o...	49	[{'comment_id': '388506910181825', 'comment_ur...	53	5157	4072	1022	55	3.0	
3	8.991520e+14	Well in, Robbo 🏆 #Euro2024 🇵🇹\n\nVâng, Robbo 🏆 ...	150	[{'comment_id': '314996431472511', 'comment_ur...	61	16019	12735	3112	146	19.0	
4	8.991290e+14	Two years ago today... Diogo 🇵🇹\n\nNgày này hai...	211	[{'comment_id': '1028198778438522', 'comment_u...	211	14403	11699	2506	125	45.0	2
...
93	8.948540e+14	"I really thank him because in this period he ...	97	[{'comment_id': '1120275918934925', 'comment_u...	43	13139	11518	1523	77	9.0	
94	8.948430e+14	Jürgen Klopp confirmed Joe Gomez and Ibrahima ...	114	[{'comment_id': '707770537958734', 'comment_ur...	34	8853	7779	934	85	3.0	
95	8.948350e+14	Anfield remembers.\n\nAnfield nhớ.	149	[{'comment_id': '7303758222967701', 'comment_u...	309	16050	13364	2481	174	8.0	1
96	8.944540e+14	Good morning 🌞 🇵🇹\n\nChào buổi sáng 🌞🇵🇹	417	[{'comment_id': '1483340398900267', 'comment_u...	124	34107	26432	7266	385	15.0	
97	8.955110e+14	Big Virg against Brentford 🇵🇹\n\nBig Virg chống...	631	[{'comment_id': '304499032539488', 'comment_ur...	727	46476	39460	6610	353	12.0	3

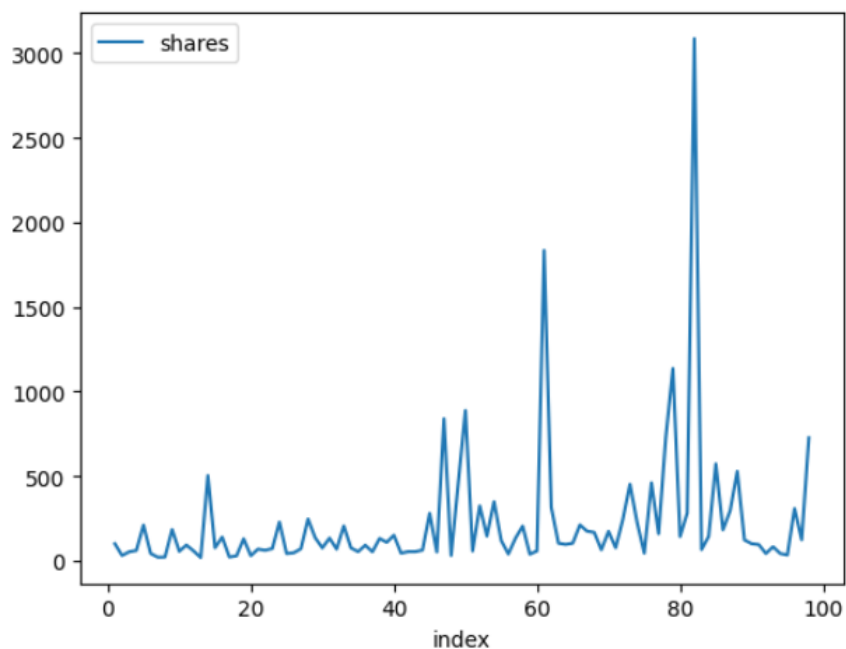
98 rows × 16 columns

Thêm cột Index

1	df = df.reset_index(inplace = False)										
2	df["index"] = df["index"] + 1										
1	df										
	index	post_id	post_text	comments	comments_full	shares	reaction_count	thích	yêu thích	thương	h
0	1	8.992690e+14	🔴👤 Emile Heskey is the latest to tell the tal...	16	[{'comment_id': '1186163892166613', 'comment_ur...	101	1175	982	183	10	
1	2	8.992110e+14	Back-to-back PFA Premier League Fans' Player o...	83	[{'comment_id': '277210241476921', 'comment_ur...	31	4594	3638	902	49	
2	3	8.990810e+14	Ryan's first goal in Red 🇵🇹\n\nA look back at o...	49	[{'comment_id': '388506910181825', 'comment_ur...	53	5157	4072	1022	55	

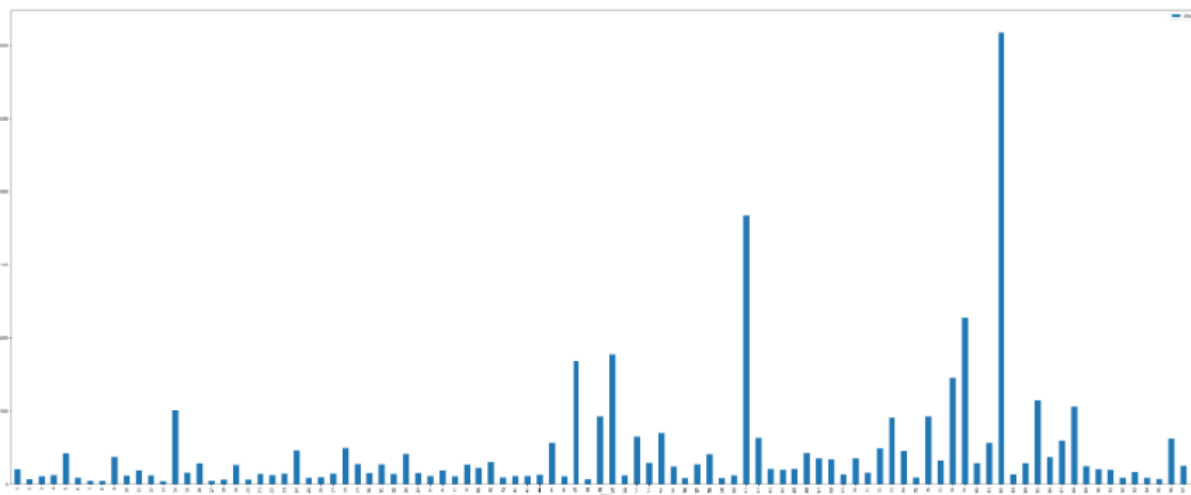
Biến động share qua từng bài đăng

```
1 df.plot(kind = "line", x = "index", y = "shares");
```



Qua biểu đồ đường trên ta thấy số lượt shares không đồng đều và đang có xu hướng tăng lên.

```
1 df.plot(kind = "bar", x = "index", y = "shares", figsize = (50, 20));
```



Qua biểu đồ trên ta thấy bài đăng thứ 82 có số share vượt trội.

Tổng số lượt share

```
1 sum(df.shares)
```

22025

Số lượt share trung bình trong 1 bài

```
1 df.shares.mean()
```

224.74489795918367

Lượt share biến động trong khoảng

```
1 df.shares.median()
```

104.0

Bài đăng có nhiều lượt share nhất

```
1 df.index[df['shares'] == max(df.shares)]
```

Index([81], dtype='int64')

```
1 df.iloc[81]
```

```
index                                82
post_id                        89512500000000.0
post_text  Are you dreaming of a Red Christmas? 🎄🌟\n\nSho...
comments                        753
comments_full  [{'comment_id': '329595236433904', 'comment_ur...
shares                        3087
reaction_count                10910
thích                        7445
yêu thích                    3258
thương thương                 105
haha                        90.0
wow                          8.0
buồn                         0.0
phấn nộ                      4.0
hour                          15
date                        11/14/2023
day                          Tuesday
Name: 81, dtype: object
```

Bài đăng có số lượt share ít nhất

```
1 df.index[df['shares'] == min(df.shares)]
```

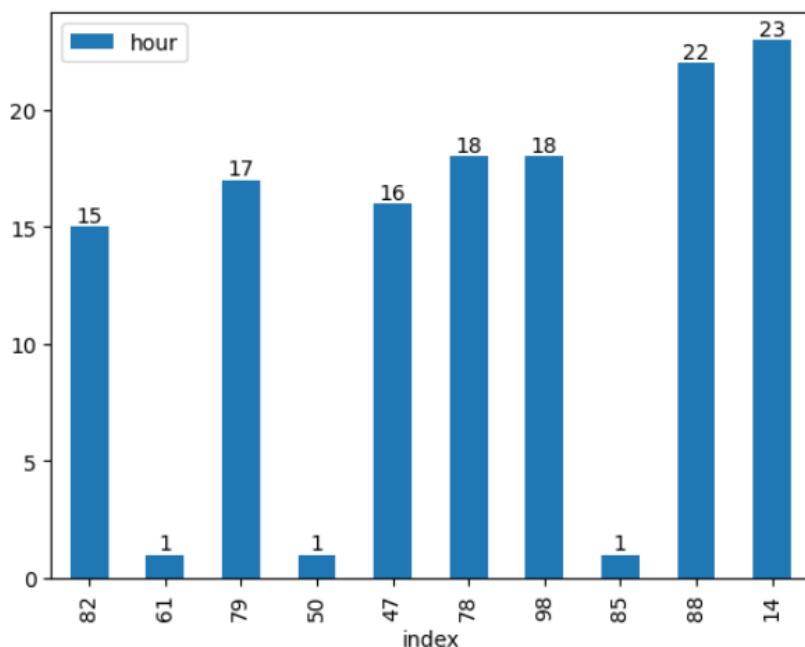
Index([12], dtype='int64')

```
1 df.iloc[12]
```

```
index                                13
post_id                        89847700000000.0
post_text    How well can you remember our 2023-24 season s...
comments                                70
comments_full    [{'comment_id': '260630993303805', 'comment_ur...
shares                                19
reaction_count                        5865
thích                                4924
yêu thích                        888
thương thương                        48
haha                                0.0
wow                                4.0
buồn                                0.0
phấn nộ                        1.0
hour                                0
date                        11/20/2023
day                        Monday
Name: 12, dtype: object
```

Top 10 bài nhiều share nhất

```
1 top_share = df.sort_values(by = ['shares'], ascending = False).head(10)
2 top_share_plt = top_share.plot(kind = "bar", x = "index", y = "hour")
3
4 for b in top_share_plt.patches:
5     top_share_plt.annotate(str(b.get_height()), (b.get_x() + b.get_width() / 2., b.get_height()),
6
7 plt.show()
```



Nội dung của top các ngày đăng bài

```
1 df.sort_values(by = ['shares'], ascending = False).head(10)[["index", "post_text"]]
```

	index	post_text
81	82	Are you dreaming of a Red Christmas? 🎄🔥\n\nSho...
60	61	Goal involvements in 1 3 consecutive Premie...
78	79	Look out, Santa 🎅\n\nCoi chừng, ông già Noel 🎅
49	50	🔥🔥🔥🔥\n\nFour goals for Mo in Egypt's 6-0 win o...
46	47	Two goals from Lucho gave Colombia a 2-1 win o...
77	78	Big Virg against Brentford 🐶\n\nBig Virg chống...
97	98	Big Virg against Brentford 🐶\n\nBig Virg chống...
84	85	The boss enjoyed our goals scored on Sunday 🍌\...
87	88	Virg with a big clearance to keep a clean shee...
13	14	Two goals in two minutes for Szobo this aftern...

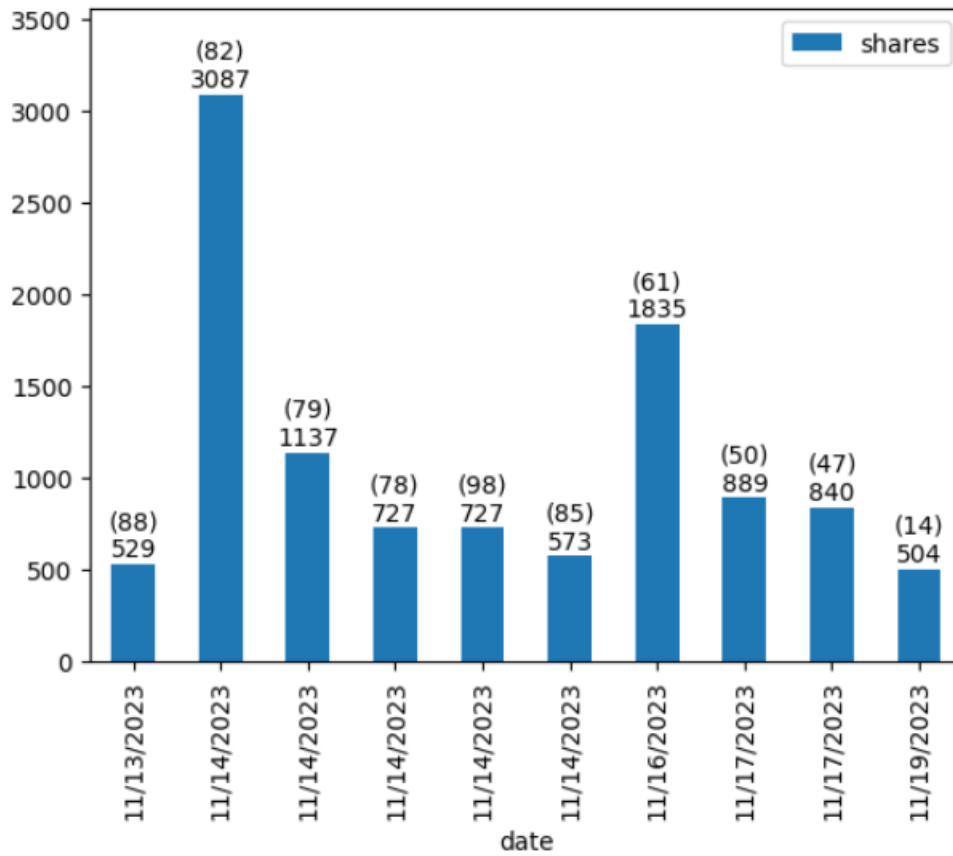
Ngày của 10 bài trong top share

```
1 df.sort_values(by = ['shares'], ascending = False).head(10)[['index', 'date']]
```

	index	date
81	82	11/14/2023
60	61	11/16/2023
78	79	11/14/2023
49	50	11/17/2023
46	47	11/17/2023
77	78	11/14/2023
97	98	11/14/2023
84	85	11/14/2023
87	88	11/13/2023
13	14	11/19/2023

Quan hệ số lượt share và ngày đăng bài trong top 10 số lượt share lớn nhất

```
1 top_share = df.sort_values(by = ['shares'], ascending = False).head(10)
2 top_share_date = top_share.sort_values(by = ['date'], ascending = True)
3 top_share_date_plt = top_share_date.plot(kind = 'bar', x = 'date', y = 'shares')
4 i = 0
5
6 for b in top_share_date_plt.patches:
7     top_share_date_plt.annotate("(" + str(top_share_date.iloc[i]["index"]) + ")")\n" + str(b.get_hei
8     i += 1
9
10 extra_space = 0.15
11 plt.ylim(top=top_share_date["shares"].max() * (1 + extra_space))
12
13 plt.show()
```



3.5 Phân tích thời gian đăng bài:

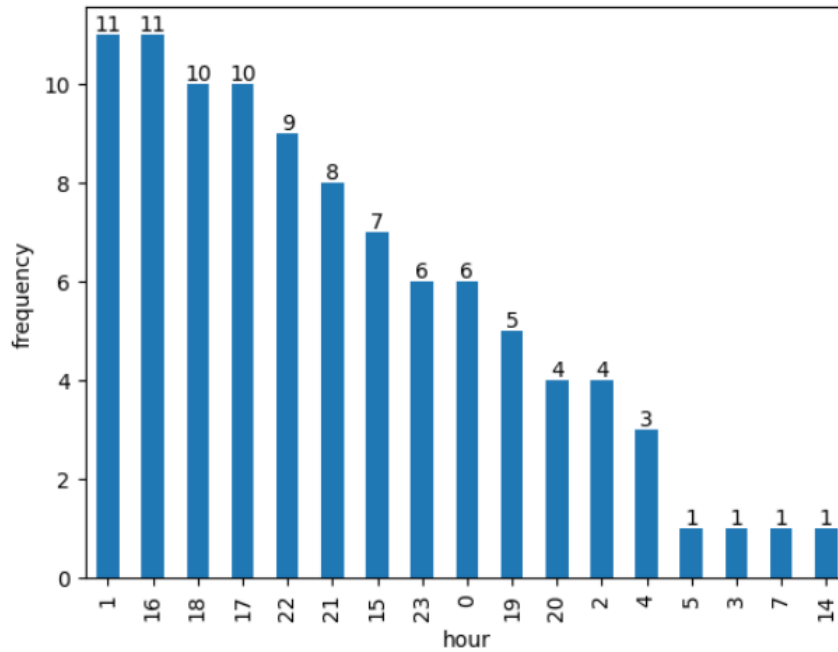
Đọc file data_ready.csv

```
1 import pandas as pd
2 import matplotlib.pyplot as plt
3 import numpy as np
```

```
1 df = pd.read_csv('C:/Users/Admin/aim/project_analysis_data_facebook/data/data_ready.csv')
```

Thời gian hay đăng bài

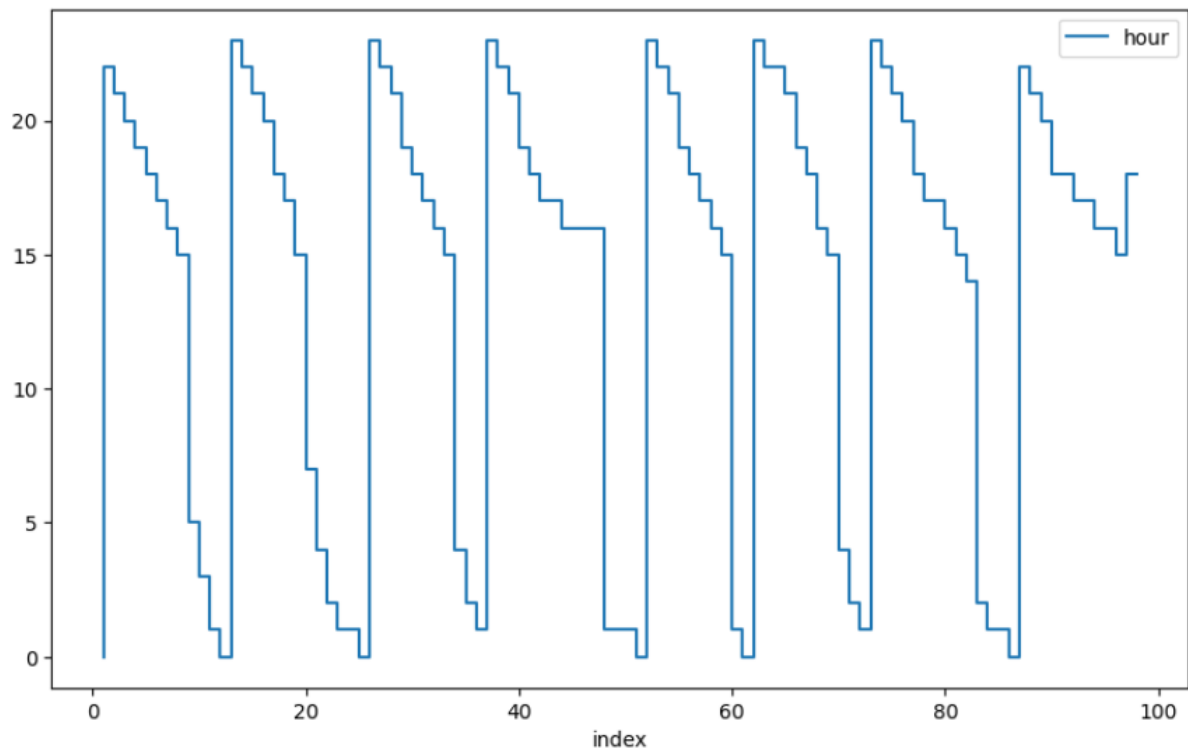
```
1 hour_plt = df.hour.value_counts().plot(kind = "bar", ylabel = 'frequency')
2
3 for b in hour_plt.patches:
4     hour_plt.annotate(b.get_height(), (b.get_x() + b.get_width() / 2, b.get_height()), ha = 'cente
5
6 plt.show()
```



Vậy admin hay chọn các giờ như 1h, 16,17,18h để đăng bài.

Biến động thời gian đăng bài

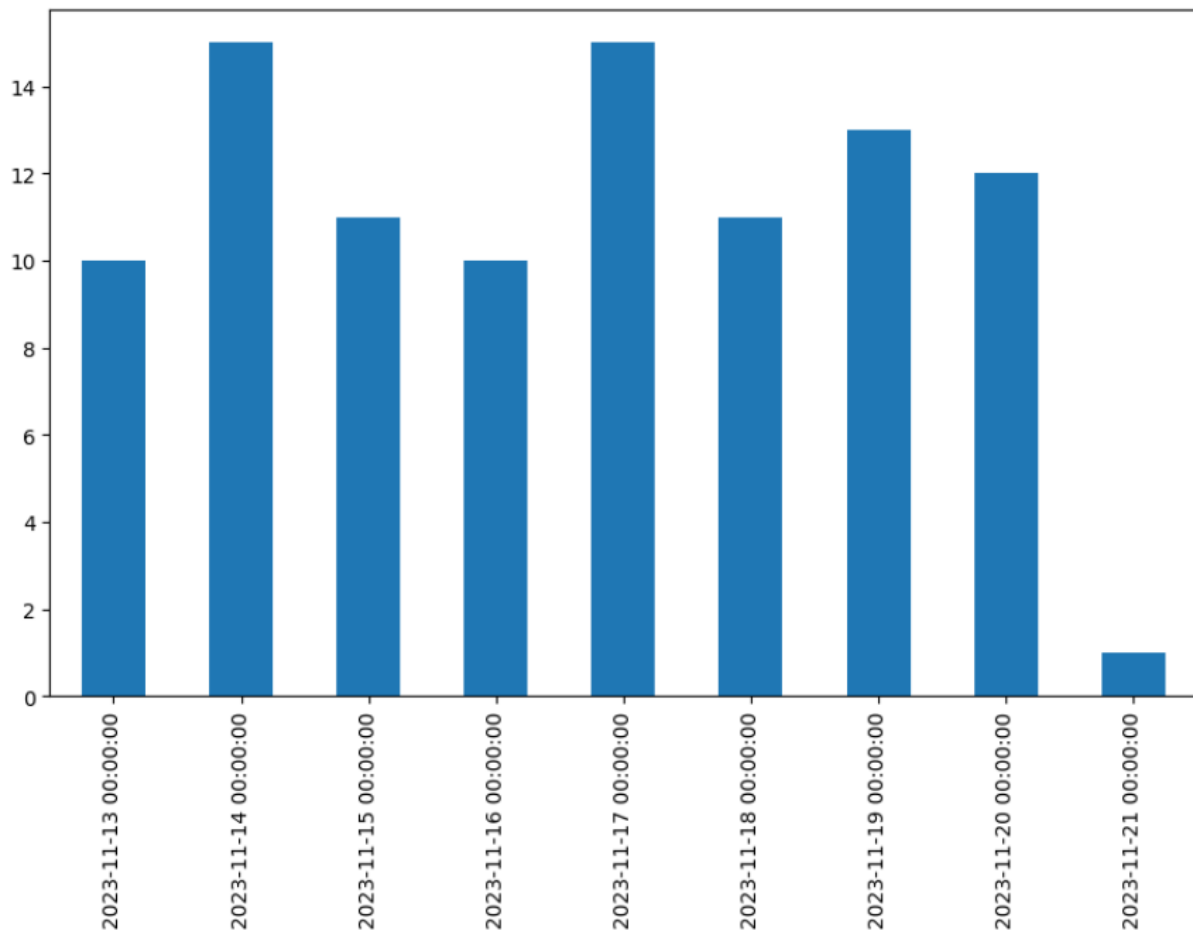
```
1 df.plot(kind = "line", x = "index", y = "hour", drawstyle='steps', figsize=(10,6));
```



Số bài đăng theo các khung giờ chênh lệch nhau quá nhiều.

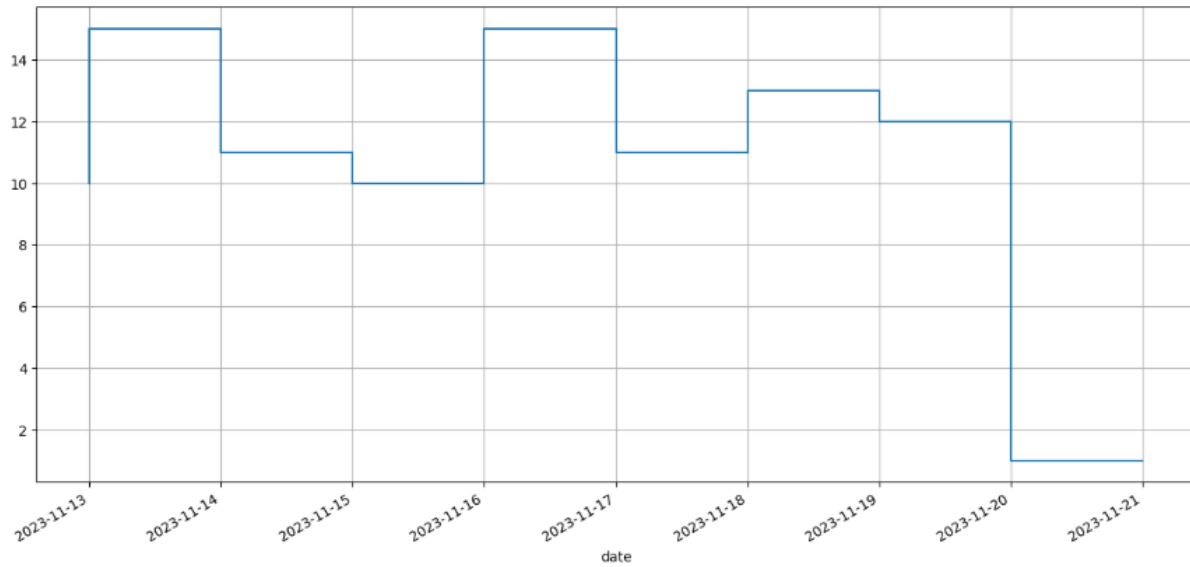
Biểu đồ biến động thời gian theo từng ngày

```
: 1 df['date'] = pd.to_datetime(df['date'])  
2 df['date'].value_counts().sort_index().plot(kind='bar', figsize=(10,6));
```



Khoảng ngày 17/11/2023 là ngày mà fanpage có nhiều bài đăng mới nhất .


```
1 df.date.value_counts().plot(kind = "line", drawstyle = "steps", grid = True, figsize = (15, 7));
```



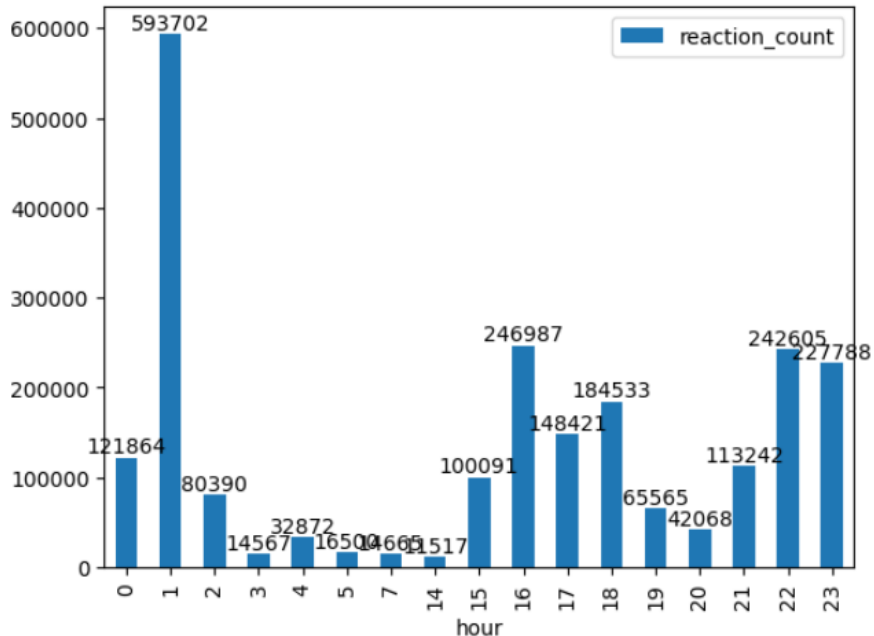
Ngày đăng bài nhiều nhất

```
1 df.date.value_counts()[df.date.value_counts() == max(df.date.value_counts())]
```

```
date
2023-11-17    15
2023-11-14    15
Name: count, dtype: int64
```

Số reaction theo giờ đăng

```
: 1 hour_react_plt = df[['hour', 'reaction_count']].groupby(['hour']).sum('reaction_count').plot(kind
2 for b in hour_react_plt.patches:
3     hour_react_plt.annotate(str(b.get_height()), (b.get_x() + b.get_width() / 2., b.get_height()),
4 plt.show()
```



Do đội bóng nằm ở nước Anh nên đa số người theo dõi nằm ở Châu Âu, vì thế cái khung 1h theo giờ Việt Nam rất được nhiều người quan tâm nên 1h là giờ mà số lượt reaction nhiều nhất.

3.6 Phân tích tương quan các trường dữ liệu:

Đọc file data_read.csv

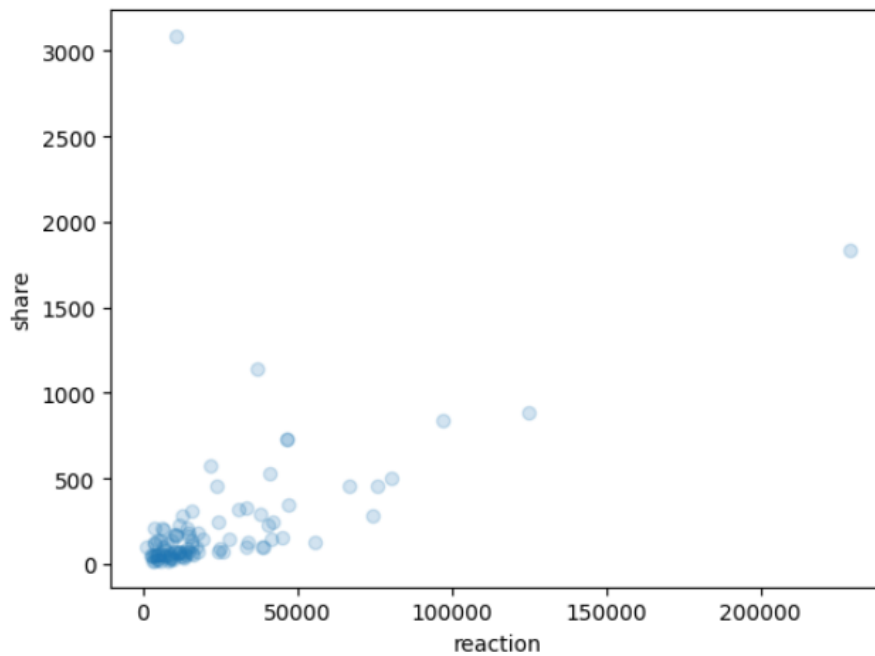
```
1 import pandas as pd
2 import matplotlib.pyplot as plt
```

```
1 df = pd.read_csv('C:/Users/Admin/aim/project_analysis_data_facebook/data/data_ready.csv')
```

```
1 df = df.reset_index(inplace = False)
2 df["index"] = df["index"] + 1
```

Sự tương quan giữa số lượng reaction và số lượt share

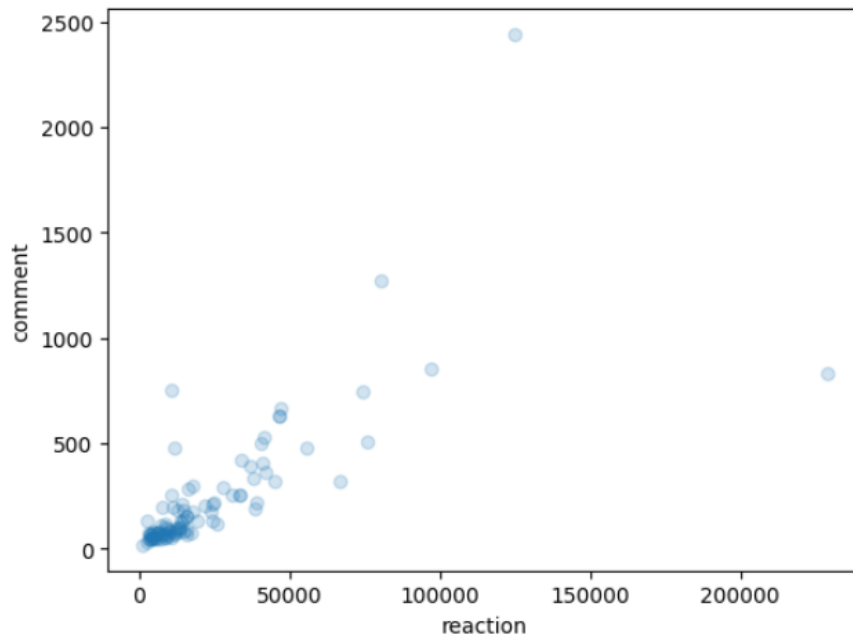
```
1 plt.scatter(df.reaction_count, df.shares, alpha = 0.2)
2
3 plt.xlabel("reaction")
4 plt.ylabel("share")
5
6 plt.show()
```



Qua biểu đồ trên ta thấy đa số các bài đăng có số lượt share không tỉ lệ thuận với số lượt reaction.

Sự tương quan giữa số lượng reaction và số lượng comment

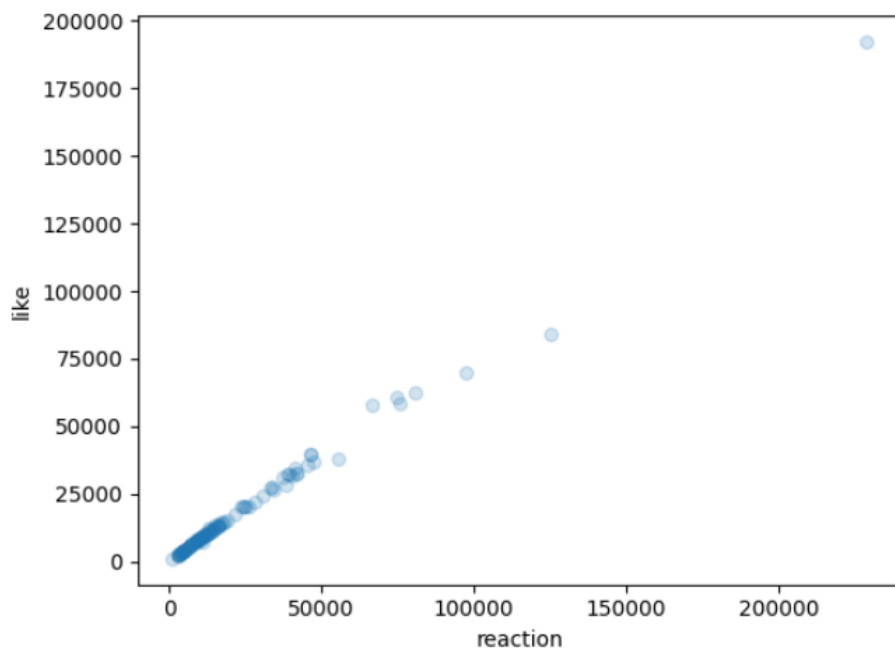
```
: 1 plt.scatter(df.reaction_count, df.comments, alpha = 0.2)
  2
  3 plt.xlabel("reaction")
  4 plt.ylabel("comment")
  5
  6 plt.show()
```



Qua biểu đồ trên ta thấy số lượt bình luận tỉ lệ thuận với số lượt thả biểu cảm.

Sự tương quan giữa số lượt biểu cảm và số lượt thích

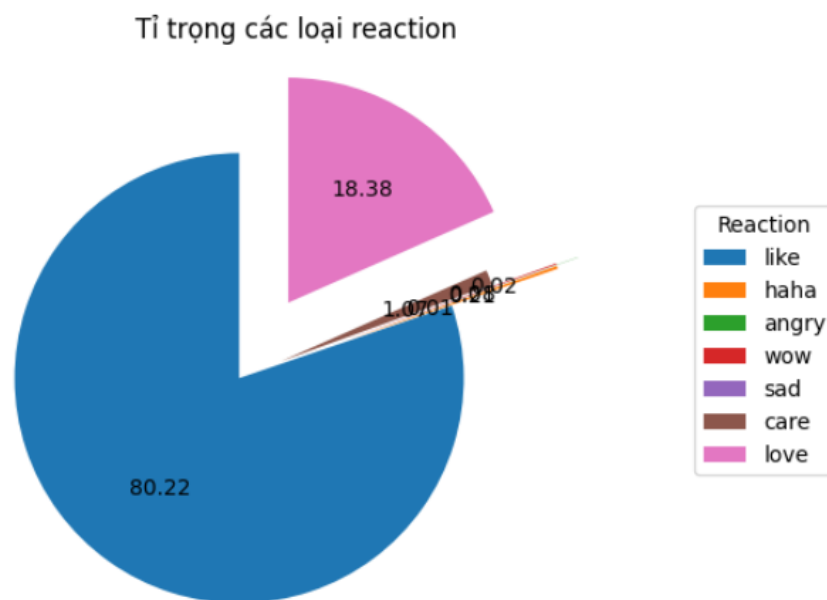
```
1 plt.scatter(df.reaction_count, df.thích, alpha = 0.2)
2
3 plt.xlabel("reaction")
4 plt.ylabel("like")
5
6 plt.show()
```



Do số lượt thích chiếm đa số trong số lượt biểu cảm nên là số lượt thích tỉ lệ thuận với số lượt reaction.

Tỉ trọng giữa các loại reaction

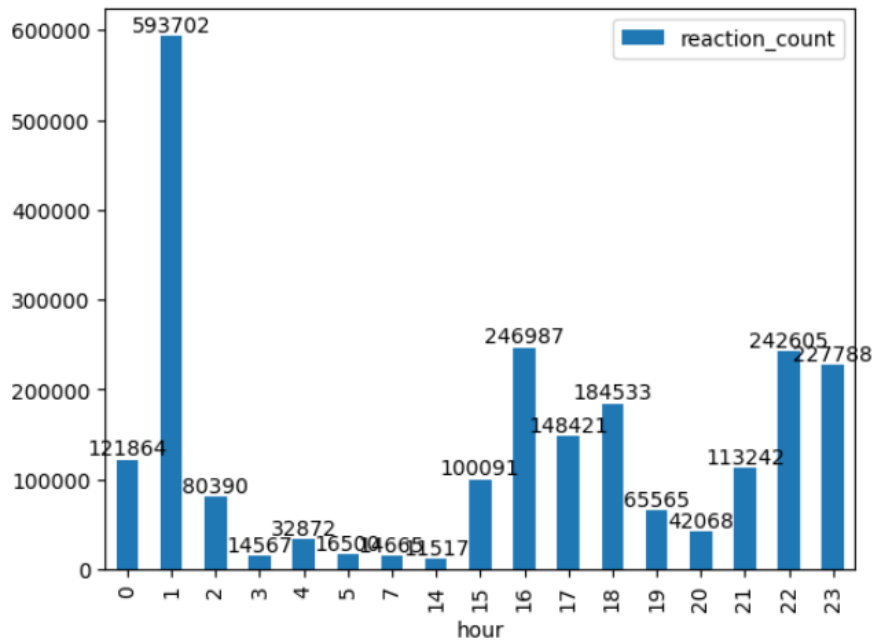
```
1 react_dict = {
2     'like': sum(df['thích']),
3     'haha': sum(df['haha']),
4     'angry': sum(df['phẫn nộ']),
5     'wow': sum(df['wow']),
6     'sad': sum(df['buồn']),
7     'care': sum(df['thương thương']),
8     'love': sum(df['yêu thích'])
9 }
10 react = []
11 number = []
12
13 explode = (0.0, 0.5, 0.6, 0.5, 0.3, 0.2, 0.4)
14
15 for x, y in react_dict.items():
16     react.append(x)
17     number.append(y)
18
19 plt.pie(number, labels=['']*len(react), autopct='%.2f', explode = explode, startangle = 90)
20
21 plt.axis('equal')
22 plt.title("Tỉ trọng các loại reaction")
23
24 plt.legend(react, title="Reaction", loc="center left", bbox_to_anchor=(1, 0.5))
25
26 plt.show()
```



Qua biểu đồ trên ta thấy số biểu cảm haha, wow, buồn, phẫn nộ, thương thương chiếm rất ít trong tỉ trọng số biểu cảm của tất cả bài đăng.

Số reaction theo giờ đăng

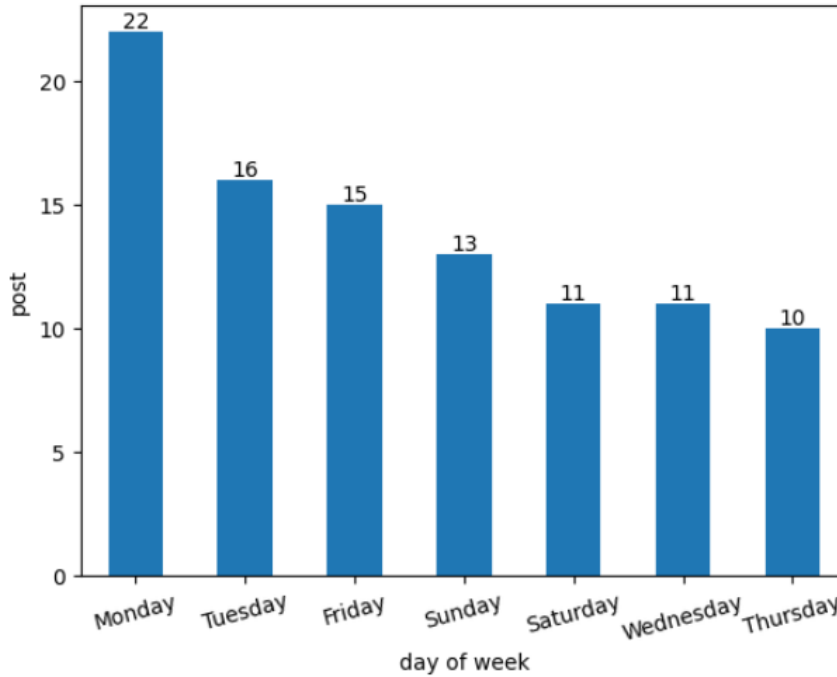
```
: 1 hour_react_plt = df[['hour', 'reaction_count']].groupby(['hour']).sum('reaction_count').plot(kind
2 for b in hour_react_plt.patches:
3     hour_react_plt.annotate(str(b.get_height()), (b.get_x() + b.get_width() / 2., b.get_height()),
4 plt.show()
```



Do đội bóng nằm ở nước Anh nên đa số người theo dõi nằm ở Châu Âu, vì thế cái khung 1h theo giờ Việt Nam rất được nhiều người quan tâm nên 1h là giờ mà số lượt reaction nhiều nhất.

Số bài đăng theo thứ

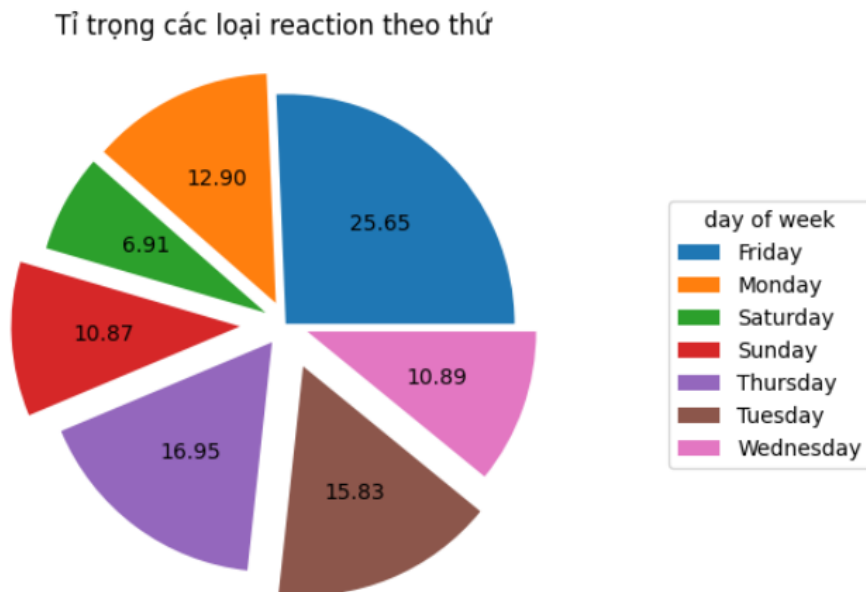
```
1 day_plt = df.day.value_counts().plot(kind = "bar", xlabel = 'day of week', ylabel = 'post', rot =  
2  
3 for b in day_plt.patches:  
4     day_plt.annotate(b.get_height(), (b.get_x() + b.get_width() / 2, b.get_height()), ha = 'center'  
5  
6 plt.show()
```



Qua biểu đồ trên ta thấy vào thứ 2 trang fanpage đăng rất nhiều bài, cũng phải thôi vì các giải bóng đá Châu Âu thường sắp xếp lịch vào chỉ nhật hàng tuần thì thứ 2 sẽ đăng rất nhiều.

Tỉ trọng reaction theo thứ

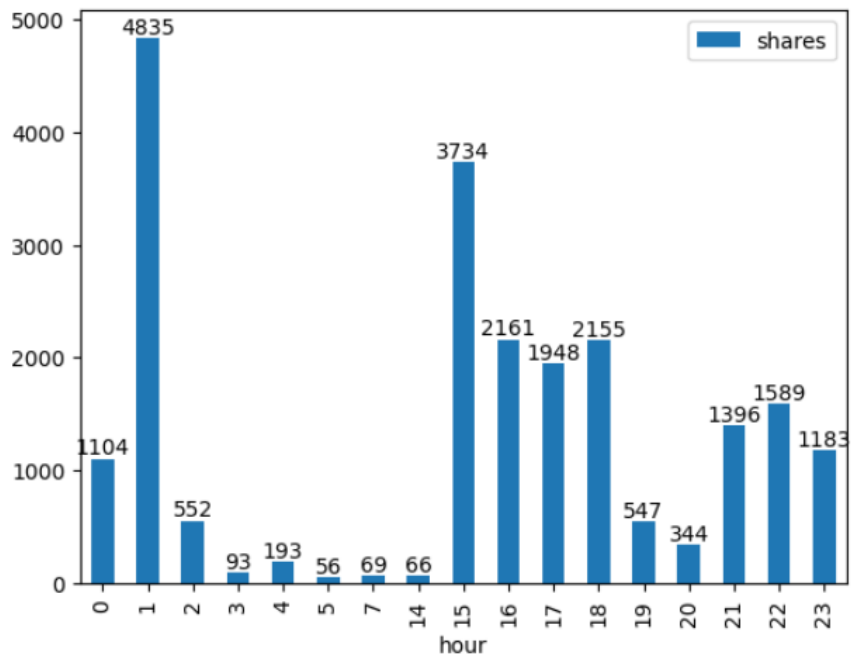
```
1 day_react = df[['day', 'reaction_count']].groupby(['day']).sum('reaction_count')
2
3 explode = (0.0, 0.1, 0.1, 0.2, 0.1, 0.2, 0.1)
4
5 plt.pie(day_react.reaction_count, labels = ['']*len(day_react.index), autopct = '%.2f', explode =
6
7 plt.axis('equal')
8 plt.title("Tỉ trọng các loại reaction theo thứ")
9
10 plt.legend(day_react.index, title="day of week", loc="center left", bbox_to_anchor=(1, 0.5))
11 plt.show()
```



Qua biểu đồ trên ta thấy vào chủ nhật thì có rất nhiều reaction vì chủ nhật mọi người được ở nhà nên thời gian họ dành thời gian cho mạng xã hội khá cao.

Tương quan giữa giờ đăng và bài share

```
1 share_hour = df[["shares", "hour"]].groupby(["hour"]).sum("shares")
2 share_hour_plt = share_hour.plot(kind = "bar")
3 for b in share_hour_plt.patches:
4     share_hour_plt.annotate(str(b.get_height()), (b.get_x() + b.get_width() / 2., b.get_height()),
5
6 plt.show()
```



3.7 Phân tích thú vị:

Code phần này nằm trong tệp analysis_interesting.


Đọc file data_ready.csv

```
1 import pandas as pd
2 import matplotlib.pyplot as plt
3 import numpy as np
```

```
1 df = pd.read_csv('C:/Users/Admin/aim/project_analysis_data_facebook/data/data_ready.csv')
```

```
1 df = df.reset_index(inplace = False)
2 df["index"] = df["index"] + 1
```

```
1 df.head()
```

	index	post_id	post_text	comments	comments_full	shares	reaction_count	thích	yêu thích	thương thương	haha	wow
0	1	8.992690e+14	 Emile Heskey is the latest to tell the tal...	16	[{'comment_id': '1186163892166613', 'comment_u...	101	1175	982	183	10	0.0	0.0
1	2	8.992110e+14	Back-to-back PFA Premier League Fans' Player o...	83	[{'comment_id': '277210241476921', 'comment_ur...	31	4594	3638	902	49	1.0	4.0
2	3	8.990810e+14	Ryan's first goal in Red 🍷\n\nA look back at o...	49	[{'comment_id': '388506910181825', 'comment_ur...	53	5157	4072	1022	55	3.0	4.0
3	4	8.991520e+14	Well in, Robbo 🍷 #Euro2024 🍷\n\nVàng, Robbo 🍷 ...	150	[{'comment_id': '314996431472511', 'comment_ur...	61	16019	12735	3112	146	19.0	6.0
4	5	8.991290e+14	Two years ago today... Diogo 🍷\n\nNgày này hai...	211	[{'comment_id': '1028198778438522', 'comment_u...	211	14403	11699	2506	125	45.0	26.0

Phân tích sự kiện

```
1 df.iloc[8].post_text
```

'A scintillating 4-0 win over Arsenal on this day in 2021 🍷'

```
1 df.iloc[25].post_text
```

'Harvey extending the England U21s lead this evening against Serbia 🍷\n\nHarvey kéo dài đội U21 Anh dẫn trước tối nay với Serbia 🍷'

```
1 df.iloc[42].post_text
```

"Diogo provided an assist for Portugal's opener in their 2-0 win over Liechtenstein 🇵🇹🇵🇹 Seleções de Portugal\n\nDiogo đã cung cấp một kiến tạo cho trận mở màn của Bồ Đào Nha trong chiến thắng 2-0 trước Liechtenstein 🇵🇹🇵🇹 Seleções de Portugal"

```
1 df.iloc[56].post_text
```

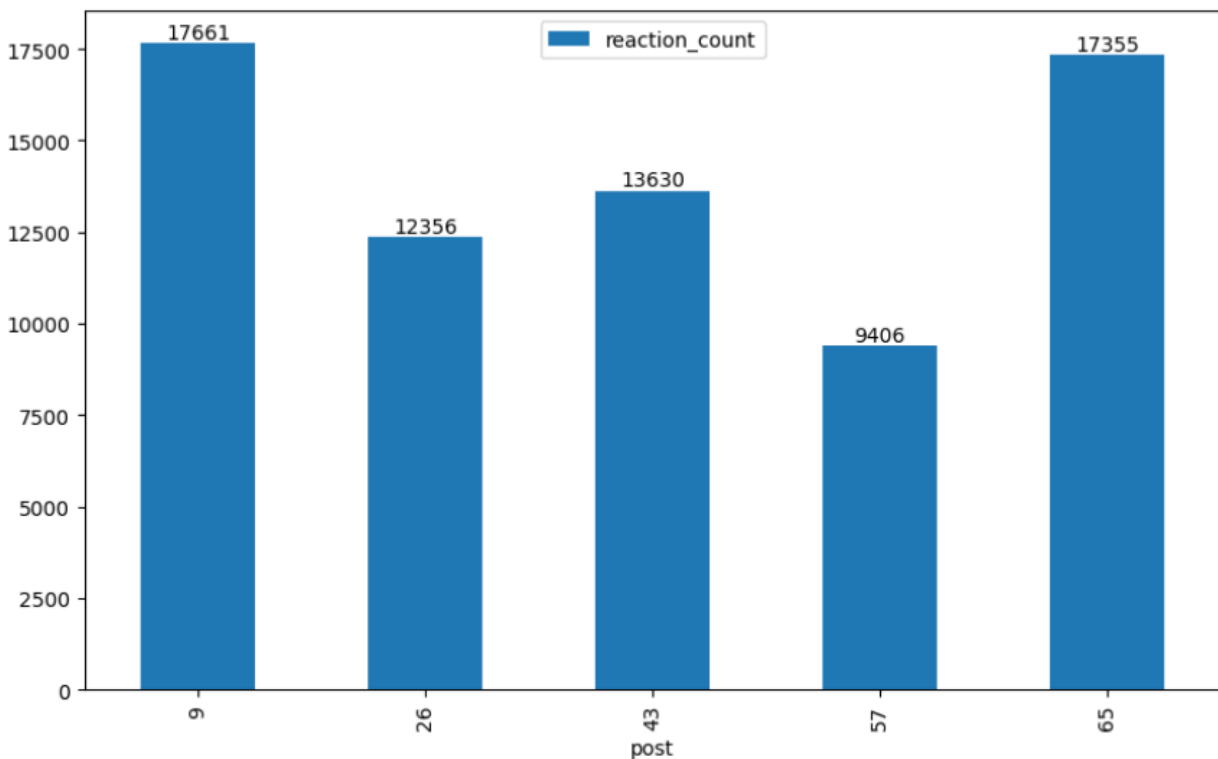
'A group of local school children were given a day to remember at the AXA Training Centre recently when they were surprised by their heroes during a Nike #GAMEON event hosted by the LFC Foundation 🙌'

```
1 df.iloc[64].post_text
```

'Looking back at our first win of the season vs Bournemouth 🏆\n\nLucho, Salah, and Diogo on target 🎯\n\nNhìn lại chiến thắng đầu tiên của chúng tôi trong mùa giải vs Bournemouth 🏆\n\nLucho, Salah, và Diogo vào mục tiêu 🎯'

```
1 top_event = df.iloc[[8,25,42,56,64]]
```

```
1 top_event_plt = top_event.plot(kind = "bar", x = "index", y = "reaction_count", xlabel = "post", f
2
3 for b in top_event_plt.patches:
4     top_event_plt.annotate(str(b.get_height()), (b.get_x() + b.get_width() / 2., b.get_height()),
5
6 plt.show()
```



Qua biểu đồ trên ta thấy khi thấy những gì liên quan đến Liverpool thắng thì mọi người thả biểu cảm rất nhiều.

Dự đoán số chia sẻ dựa trên mô hình học máy

Khai báo thư viện

```
1 import pandas as pd
2 from sklearn.model_selection import train_test_split
3 from sklearn.ensemble import GradientBoostingRegressor
4 from sklearn.metrics import mean_squared_error
5 import matplotlib.pyplot as plt
```

Load dữ liệu từ file CSV hoặc nguồn dữ liệu khác

```
1 data = pd.read_csv('C:/Users/Admin/aim/project_analysis_data_facebook/data/data_ready.csv')
```

Chọn các đặc trưng và mục tiêu

```
1 features = ['comments', 'thích', 'haha', 'yêu thích', 'wow', 'buồn', 'phẫn nộ', 'thương thương']
2 target = 'shares'
3
4 X = data[features]
5 y = data[target]
```

Chia dữ liệu thành tập huấn luyện và tập kiểm tra

```
1 X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Khởi tạo mô hình Gradient Boosting Regressor

```
1 model = GradientBoostingRegressor(n_estimators=100, learning_rate=0.1, max_depth=3, random_state=42)
```

Huân luyện trên tập huấn luyện

```
1 model.fit(X_train, y_train)
```

```
c:\Users\Admin\AppData\Local\Programs\Python\Python39\lib\site-packages\sklearn\utils\validation.py:767: FutureWarning: is_sparse is deprecated and will be removed in a future version. Check `isinstance(dtype, pd.SparseDtype)` instead.
    if not hasattr(array, "sparse") and array.dtypes.apply(is_sparse).any():
c:\Users\Admin\AppData\Local\Programs\Python\Python39\lib\site-packages\sklearn\utils\validation.py:605: FutureWarning: is_sparse is deprecated and will be removed in a future version. Check `isinstance(dtype, pd.SparseDtype)` instead.
    if is_sparse(pd_dtype):
c:\Users\Admin\AppData\Local\Programs\Python\Python39\lib\site-packages\sklearn\utils\validation.py:614: FutureWarning: is_sparse is deprecated and will be removed in a future version. Check `isinstance(dtype, pd.SparseDtype)` instead.
    if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
c:\Users\Admin\AppData\Local\Programs\Python\Python39\lib\site-packages\sklearn\utils\validation.py:605: FutureWarning: is_sparse is deprecated and will be removed in a future version. Check `isinstance(dtype, pd.SparseDtype)` instead.
    if is_sparse(pd_dtype):
c:\Users\Admin\AppData\Local\Programs\Python\Python39\lib\site-packages\sklearn\utils\validation.py:614: FutureWarning: is_sparse is deprecated and will be removed in a future version. Check `isinstance(dtype, pd.SparseDtype)` instead.
    if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
```

GradientBoostingRegressor(random_state=42)

In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.

On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

Dự đoán trên tập kiểm tra

```
1 y_pred = model.predict(X_test)
2
```

```
c:\Users\Admin\AppData\Local\Programs\Python\Python39\lib\site-packages\sklearn\utils\validation.py:767: FutureWarning: is_sparse is deprecated and will be removed in a future version. Check `isinstance(dtype, pd.SparseDtype)` instead.
    if not hasattr(array, "sparse") and array.dtypes.apply(is_sparse).any():
c:\Users\Admin\AppData\Local\Programs\Python\Python39\lib\site-packages\sklearn\utils\validation.py:605: FutureWarning: is_sparse is deprecated and will be removed in a future version. Check `isinstance(dtype, pd.SparseDtype)` instead.
    if is_sparse(pd_dtype):
c:\Users\Admin\AppData\Local\Programs\Python\Python39\lib\site-packages\sklearn\utils\validation.py:614: FutureWarning: is_sparse is deprecated and will be removed in a future version. Check `isinstance(dtype, pd.SparseDtype)` instead.
    if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
```

Đánh giá hiệu suất của mô hình

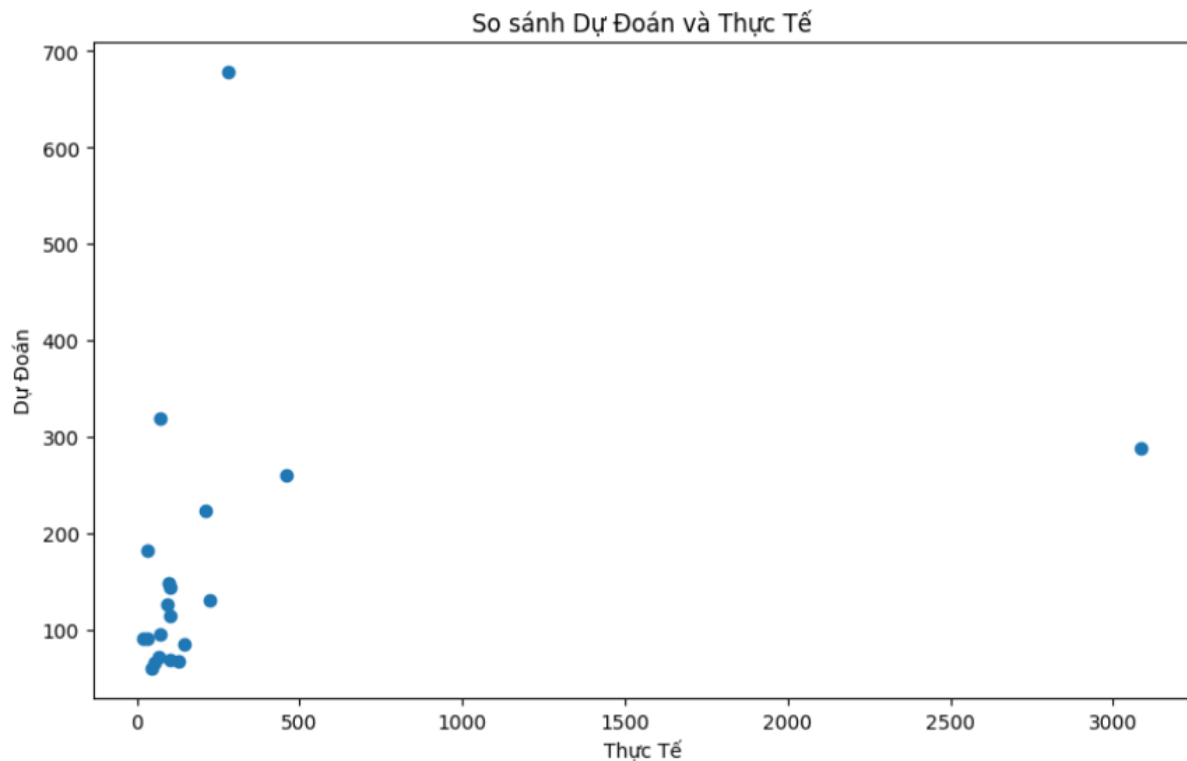
```
1 mse = mean_squared_error(y_test, y_pred)
2 print(f'Mean Squared Error: {mse}')
```

Mean Squared Error: 407592.8321141412

```
c:\Users\Admin\AppData\Local\Programs\Python\Python39\lib\site-packages\sklearn\utils\validation.py:605: FutureWarning: is_sparse is deprecated and will be removed in a future version. Check `isinstance(dtype, pd.SparseDtype)` instead.
    if is_sparse(pd_dtype):
c:\Users\Admin\AppData\Local\Programs\Python\Python39\lib\site-packages\sklearn\utils\validation.py:614: FutureWarning: is_sparse is deprecated and will be removed in a future version. Check `isinstance(dtype, pd.SparseDtype)` instead.
    if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
```

Trực quan hóa dự đoán và kết quả thực tế

```
1 plt.figure(figsize=(10,6))
2 plt.scatter(y_test, y_pred)
3 plt.xlabel('Thực Tế')
4 plt.ylabel('Dự Đoán')
5 plt.title('So sánh Dự Đoán và Thực Tế')
6 plt.show()
```



Qua biểu đồ ta thấy kết quả dự đoán số lượt share => các dữ liệu về trường share không đồng đều hoặc tập dữ liệu hơi ít.