**ĐẠI HỌC QUỐC GIA HÀ NỘI**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ**



**Họ và tên: Nguyễn Tuấn Thành**

**Mã sinh viên: 22022624**

**Đồ án môn: Lập trình xử lí 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**

**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 dể báo cáo của em dược 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. File [www.facebook.cookies.txt](http://www.facebook.cookies.txt) ................................................................ 5

1.3Thu 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

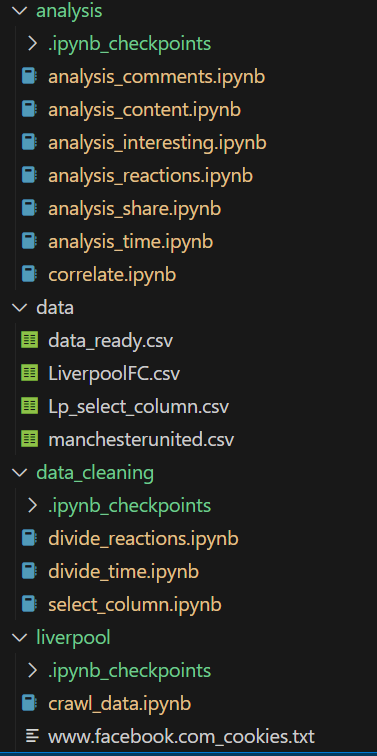
3.5 Phân tích thời gian đăng bài ................................................................. 37

3.6 Phân tích **tương quan** các trường dữ liệu ............................................ 42

3.7 Phân tích **thú vị** về dữ liệu .................................................................... 51

3.8 So sánh với Fanpage Manchester United ............................................. 55

Các tệp có khung sau:

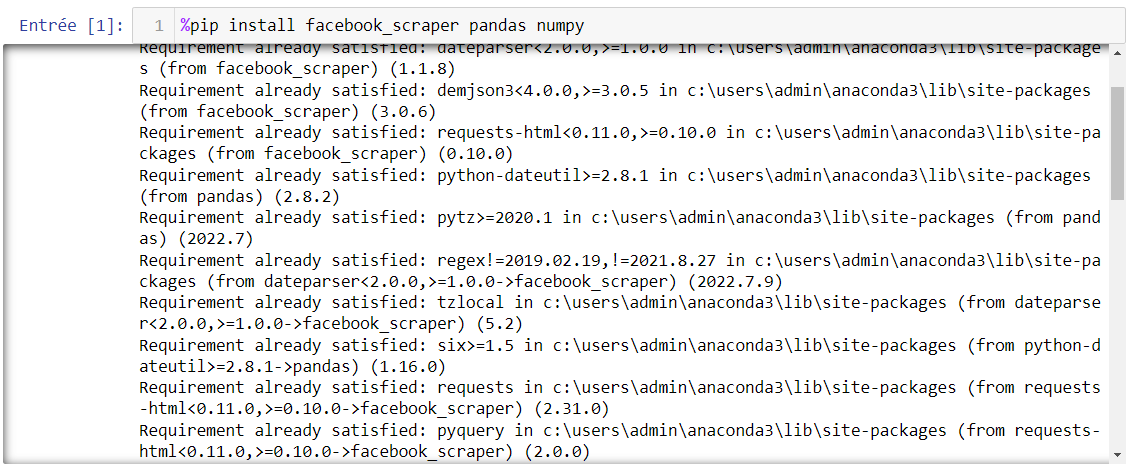


Link GitHub: <https://github.com/TuanThanh2004/project_analysis_data_facebook.git>

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

* 1. 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.



1.2 File [www.facebook.cookies.txt](http://www.facebook.cookies.txt):

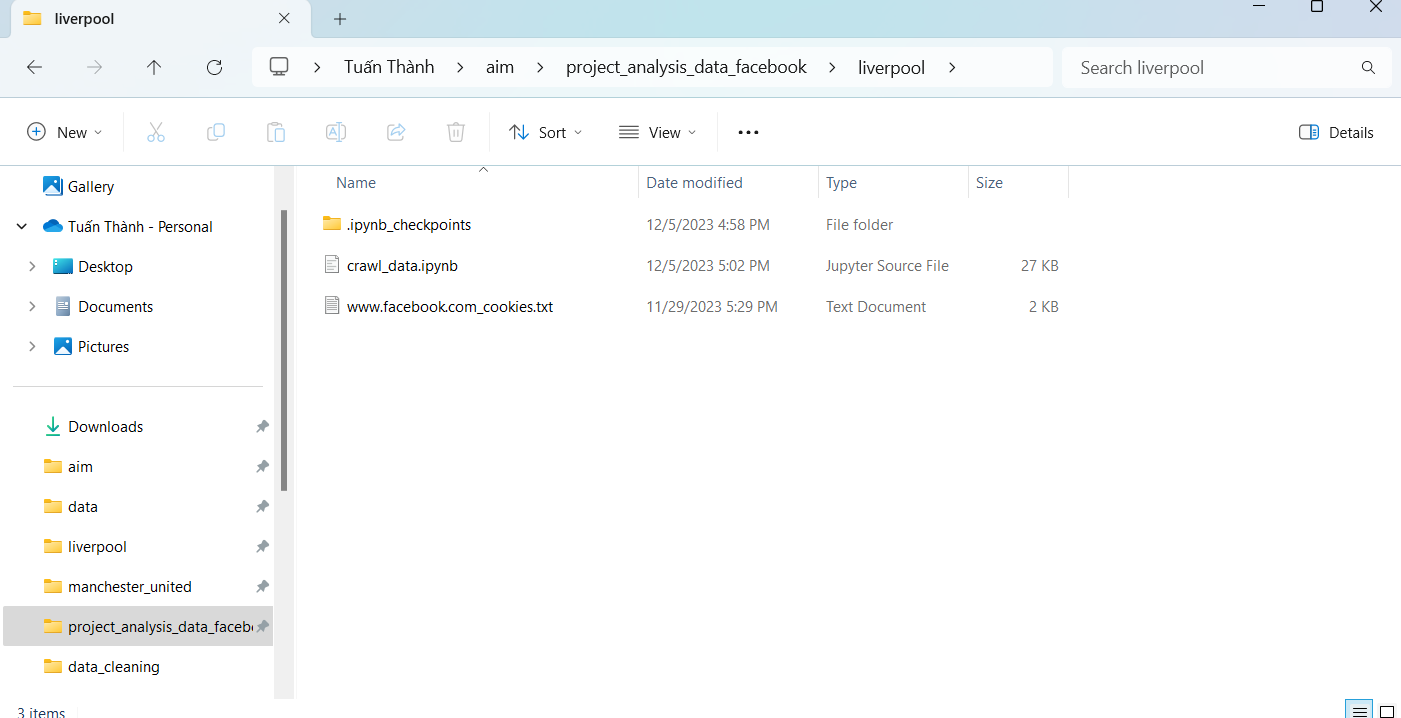
Dùng extension Get cookies.txt trên chrome



Nhấn Expost và lưu vào tệp data



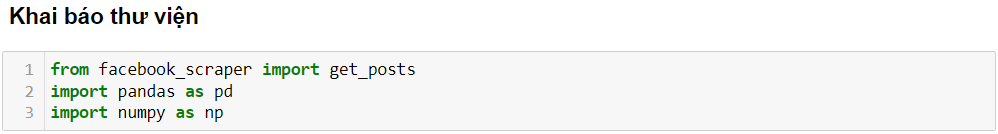
Lưu cookies vào trong thư mục có tên đội bóng.

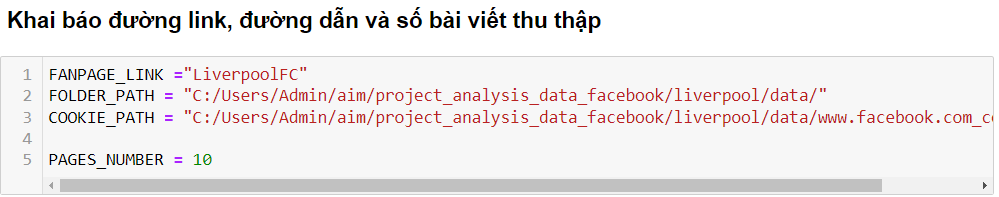


1.3Thu thập dữ liệu về:

Dùng thư viện facebook\_scraper.

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



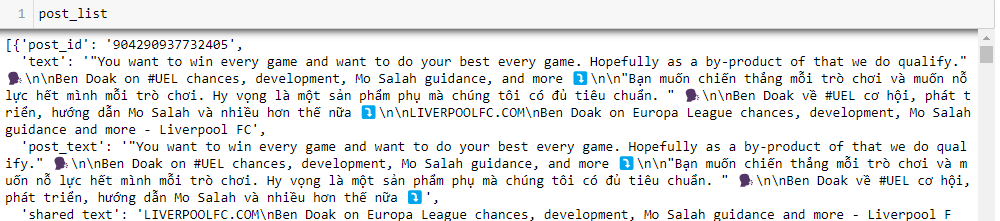


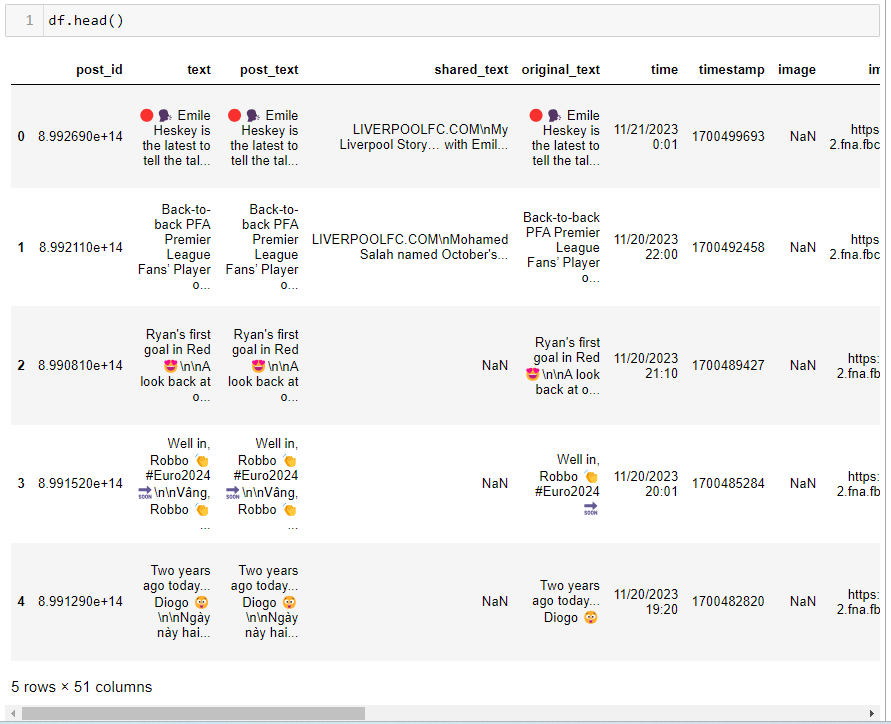


1.4 Lưu liệu thu thập được:



1.5 Xem thông tin dữ liệu vừa thu thập được:





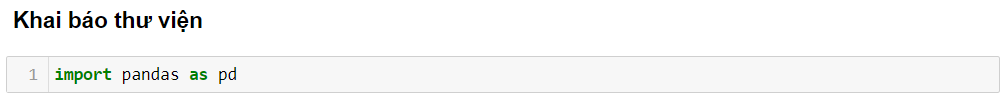


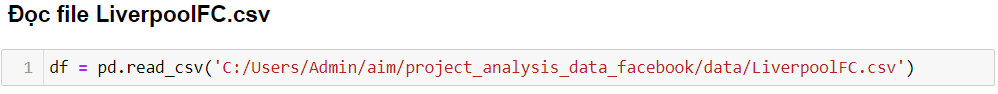
Phần 2 : Làm sạch và xử lí dữ liệu:

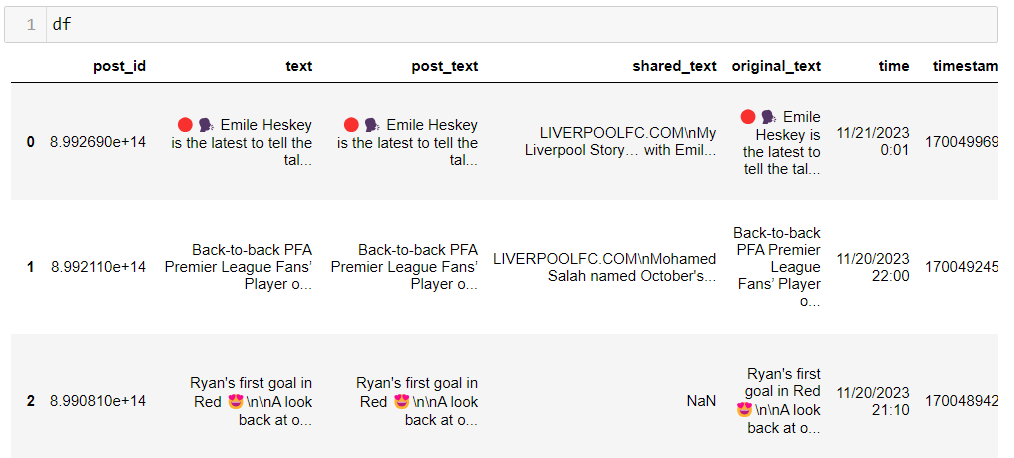
Tất cả file liên quan phần này nằm trong thực data\_cleaning.

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

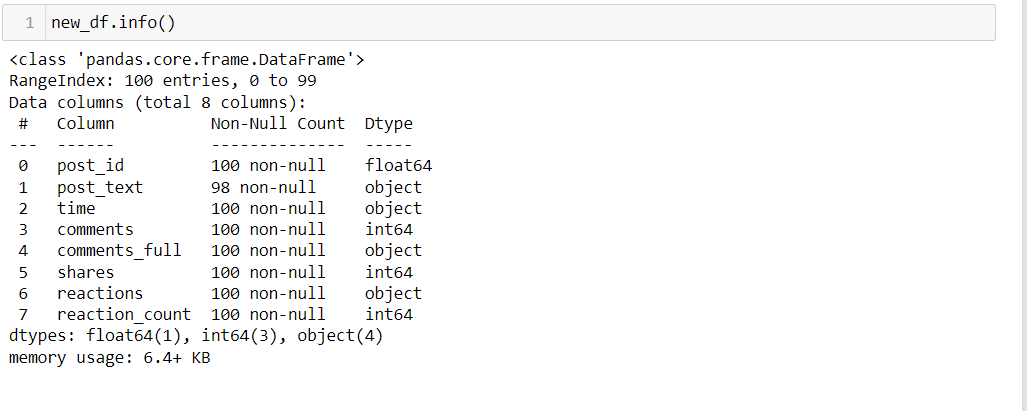








**new\_df = df[["post\_id", "post\_text", "time", "comments","comments\_full", "shares", "reactions", "reaction\_count"]]**



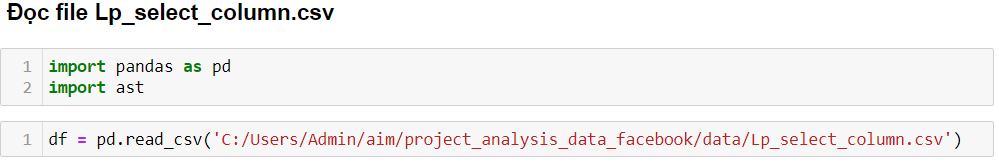


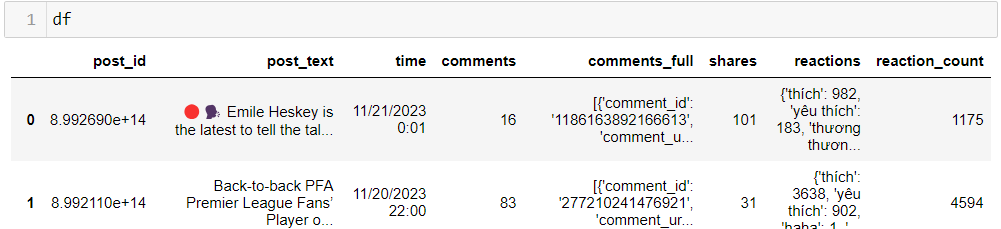


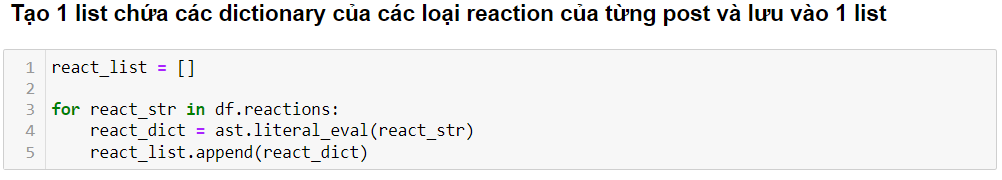
new\_df\_01.to\_csv('C:/Users/Admin/aim/project\_analysis\_data\_facebook/data/Lp\_select\_column.csv', index = False)

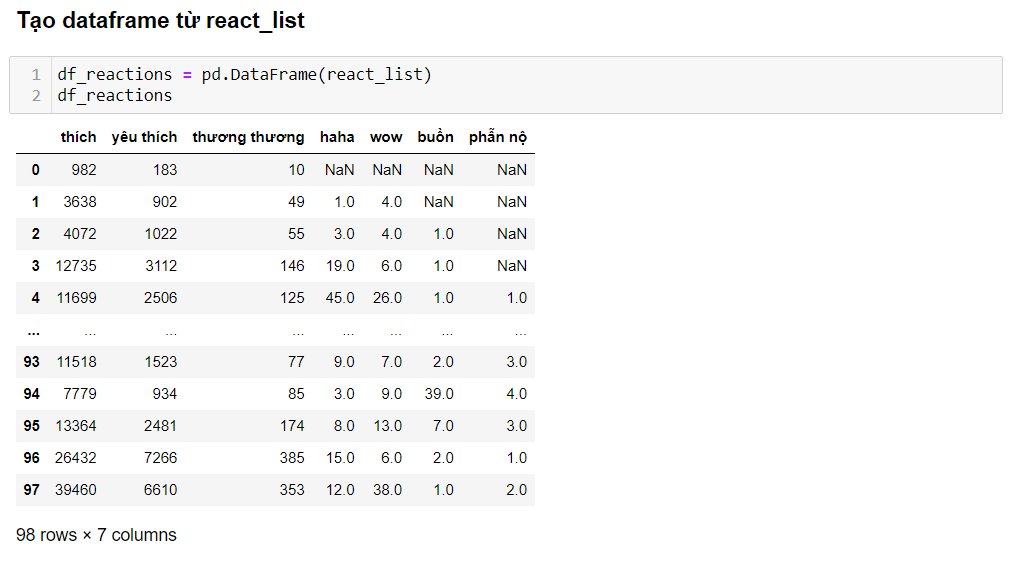
2.2 Tách cột biểu cảm thành các trường cần phân tích:

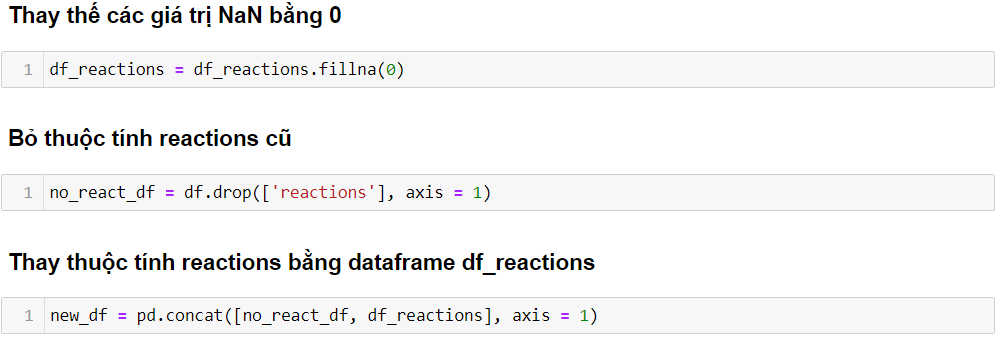
code này được lưu trong thư mục data\_cleaninng tệp tên là divide\_reactions.ipynb

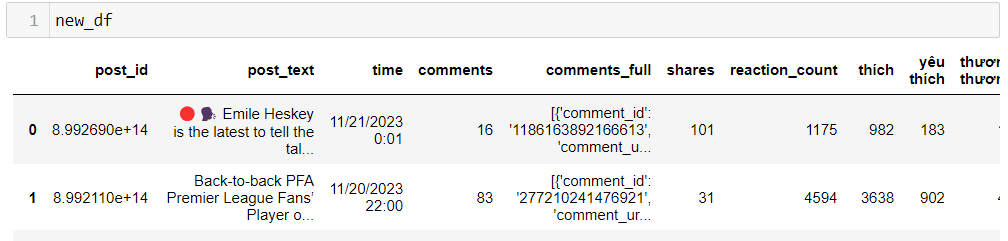


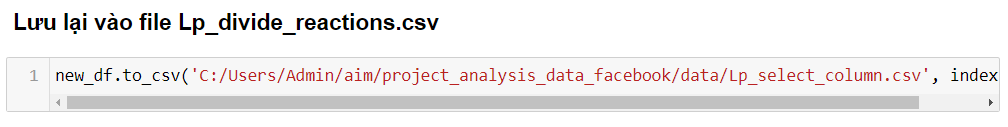






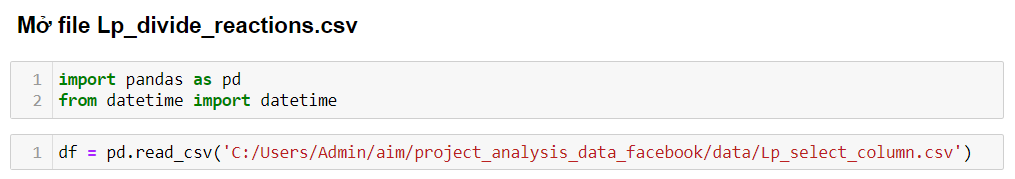


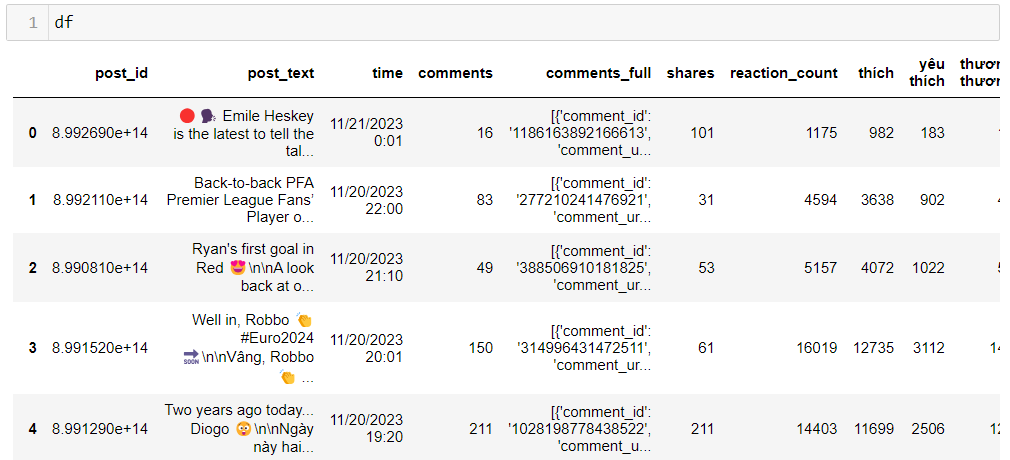


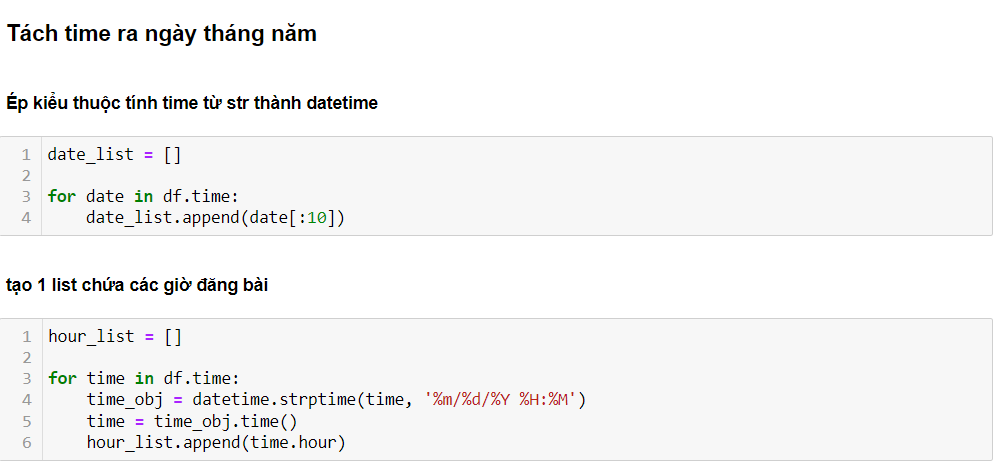


2.3 Tách cột thời gian thành cột ngày tháng năm , thứ , giờ:

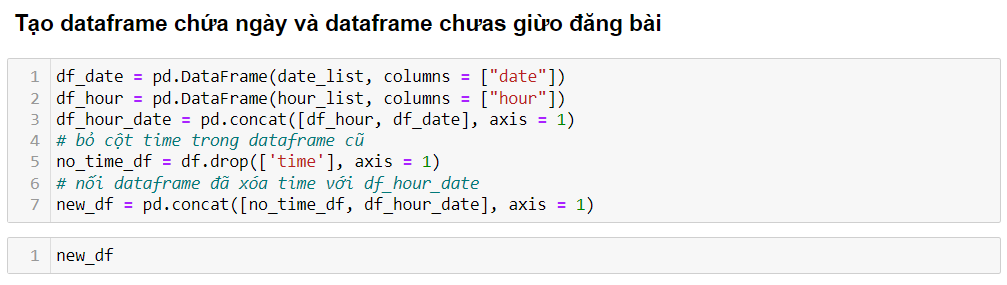
Code của phần này nằm trong tệp divide\_time.ipynb của thư mục data\_cleaning

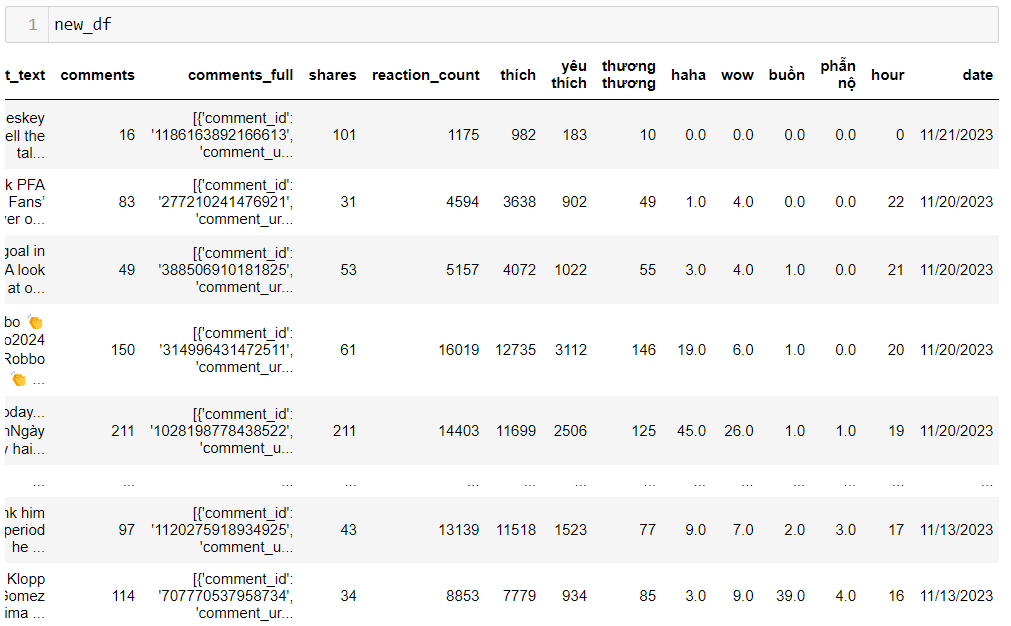


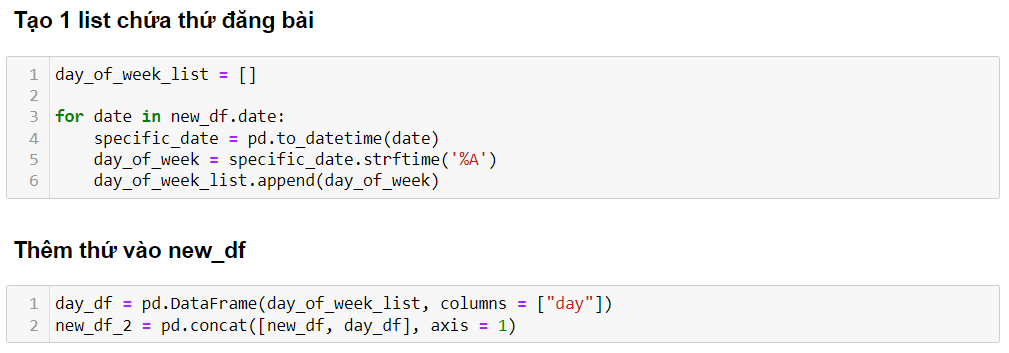


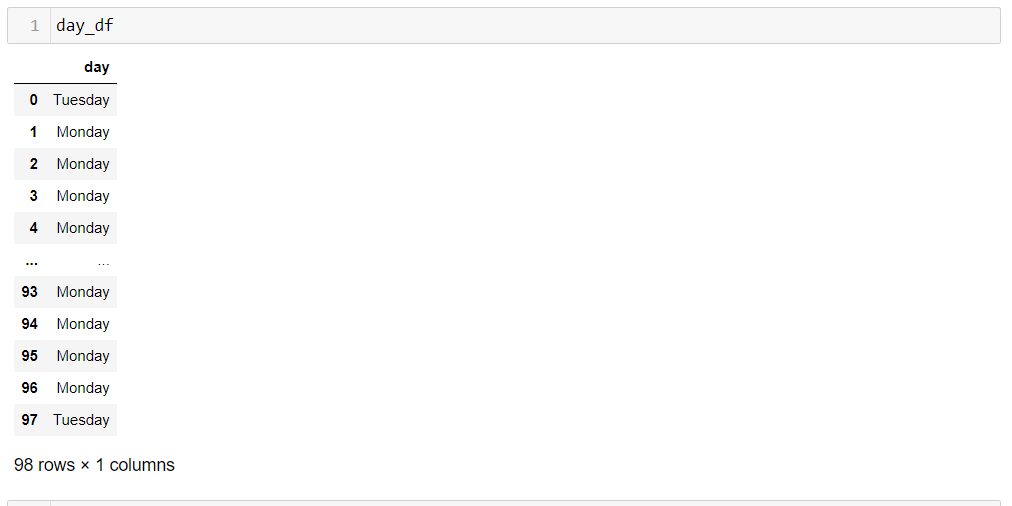


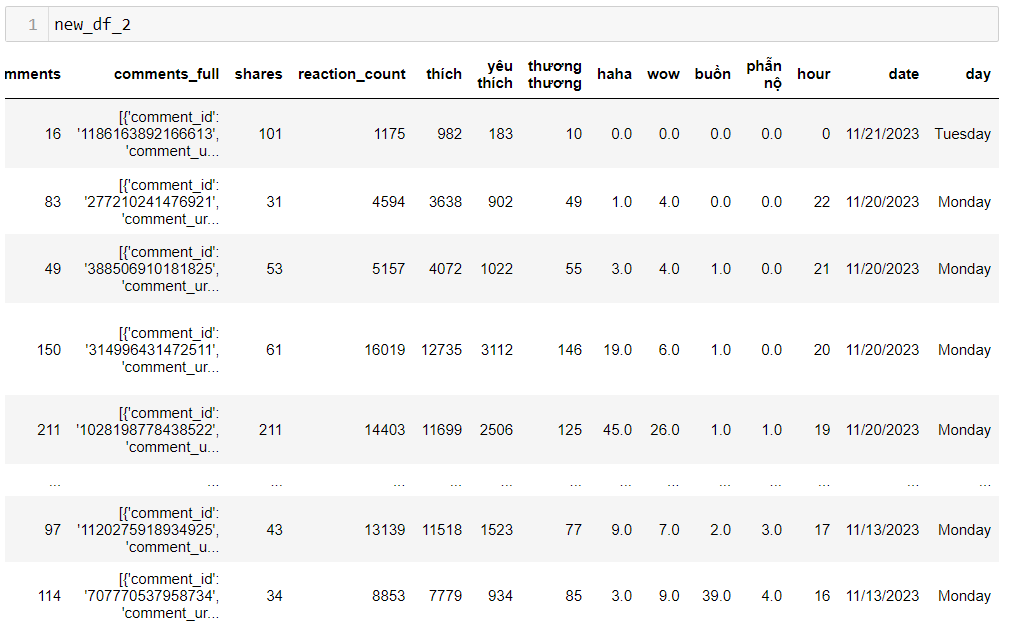


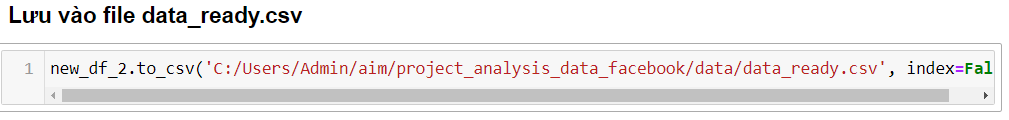










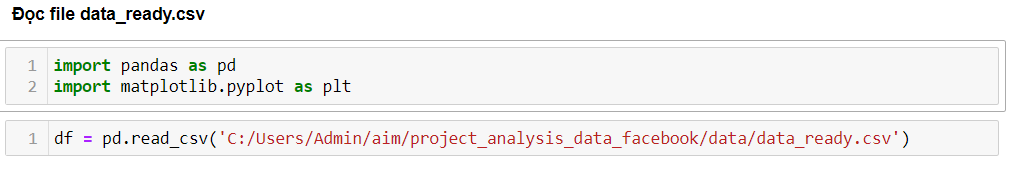


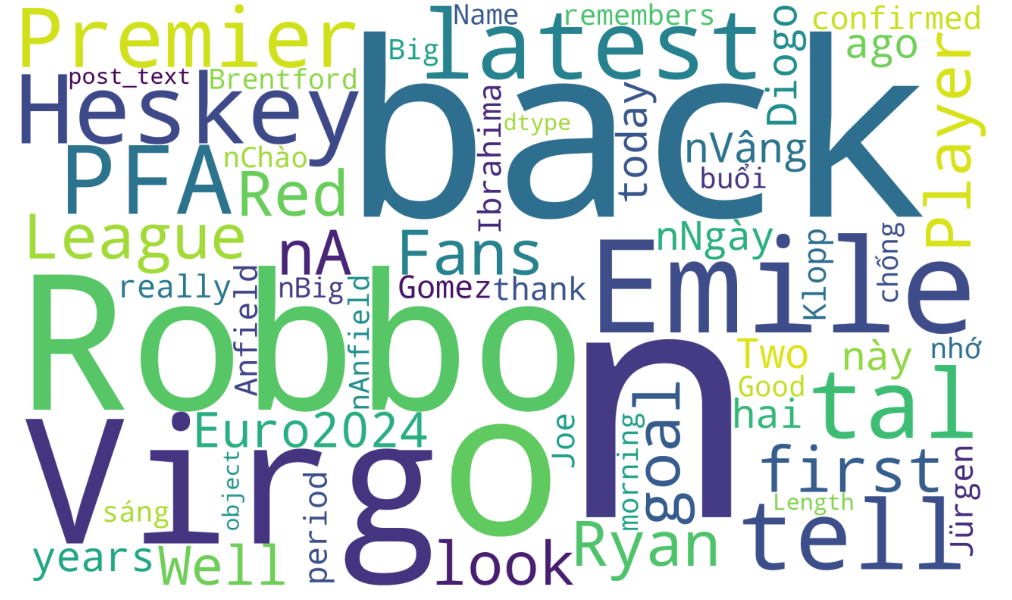
Phần 3 : Phân tích dữ liệu:

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



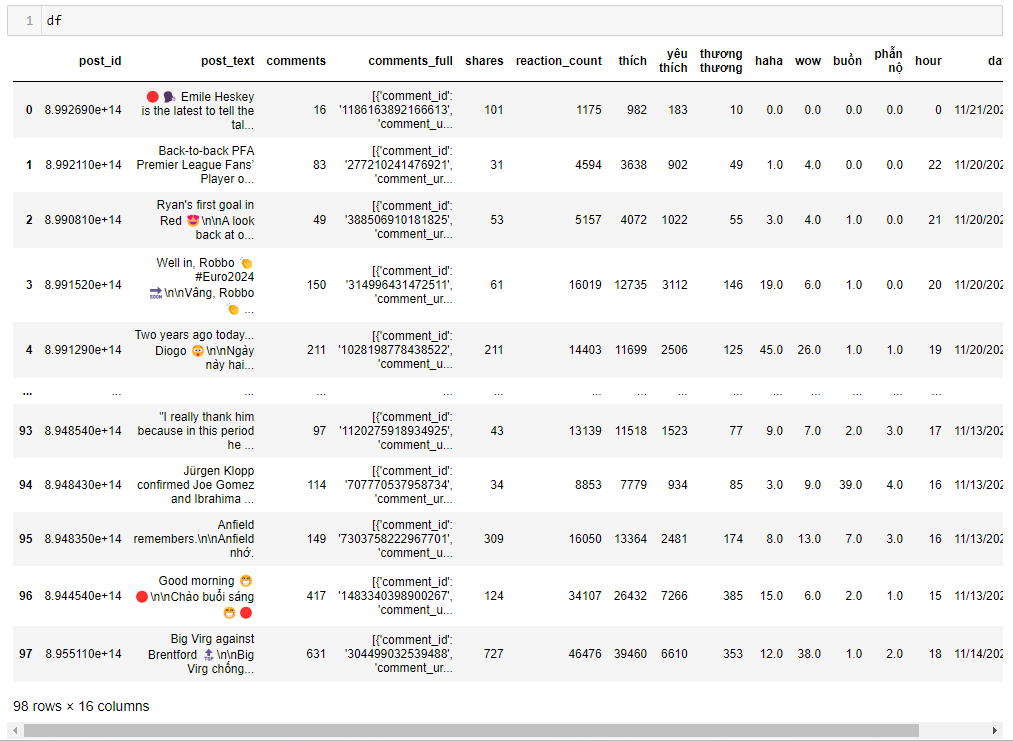
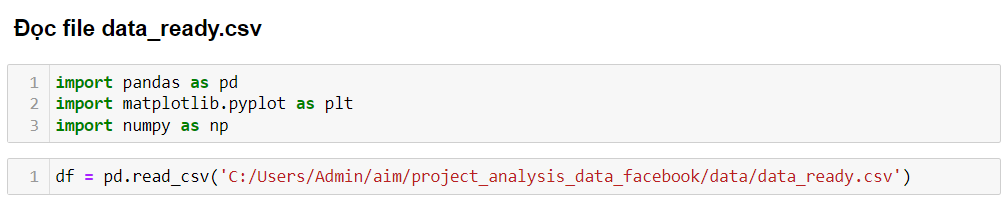


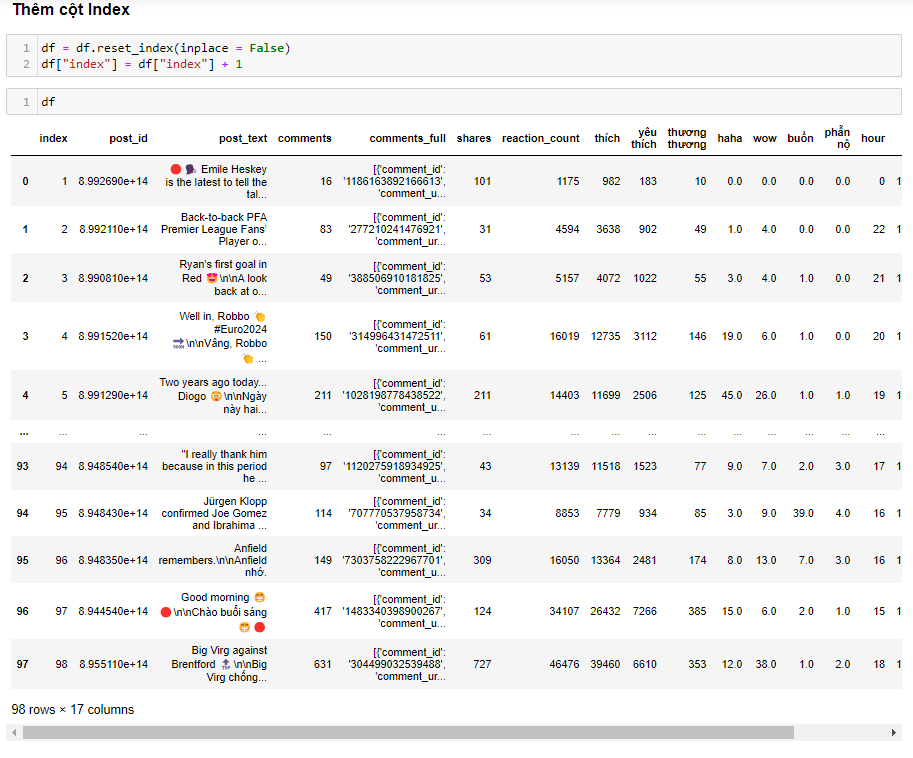
Nhìn biểu đồ trên ta thấy từ back xuất hiện nhiều nhất.

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

3.2 Phân tích bình luận:

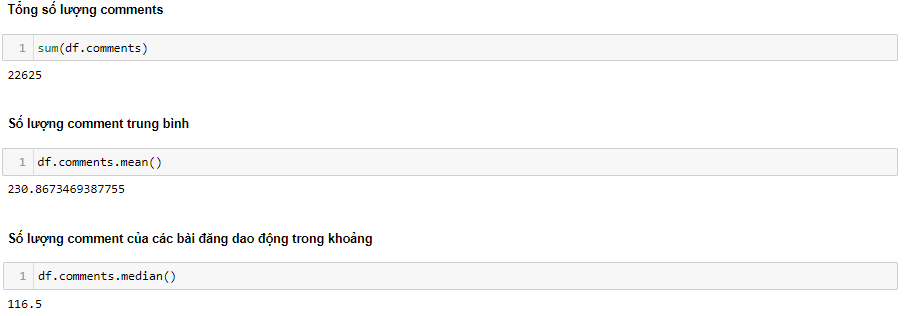
Code phần này nằm trong tệp analysis\_comments.ipynb



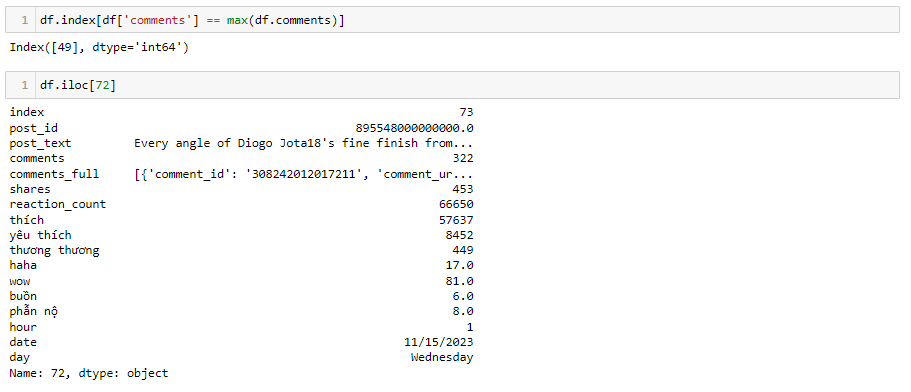




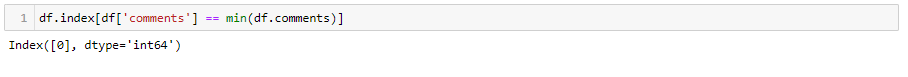
Kết luận: Bài viết thứ 50 là bài viết có số lượt commen nhiều nhất.



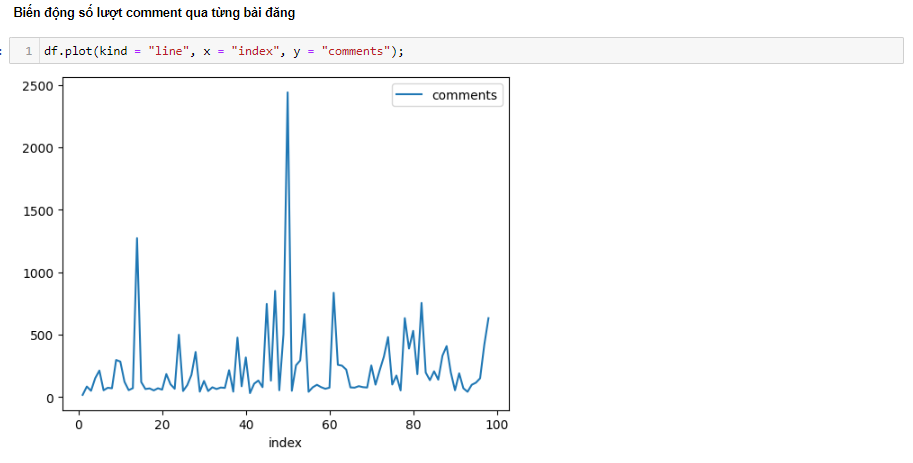
Bài viết có số lượt comment lớn nhất



Bài viết có số comments nhỏ nhất



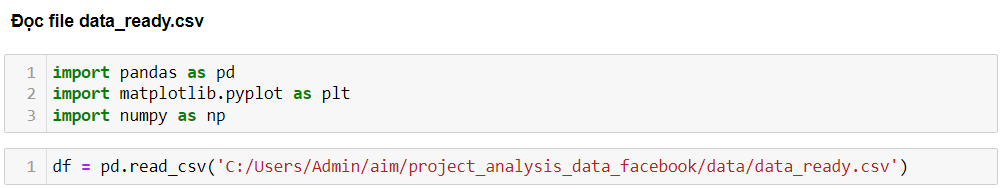
Biểu đồ đường thể hiện sự biến động của trường comments:

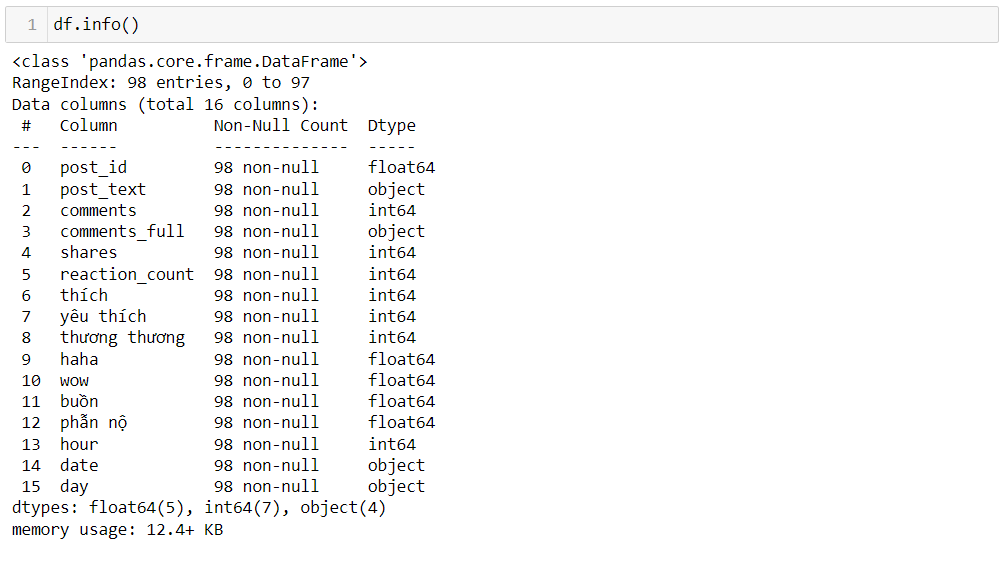


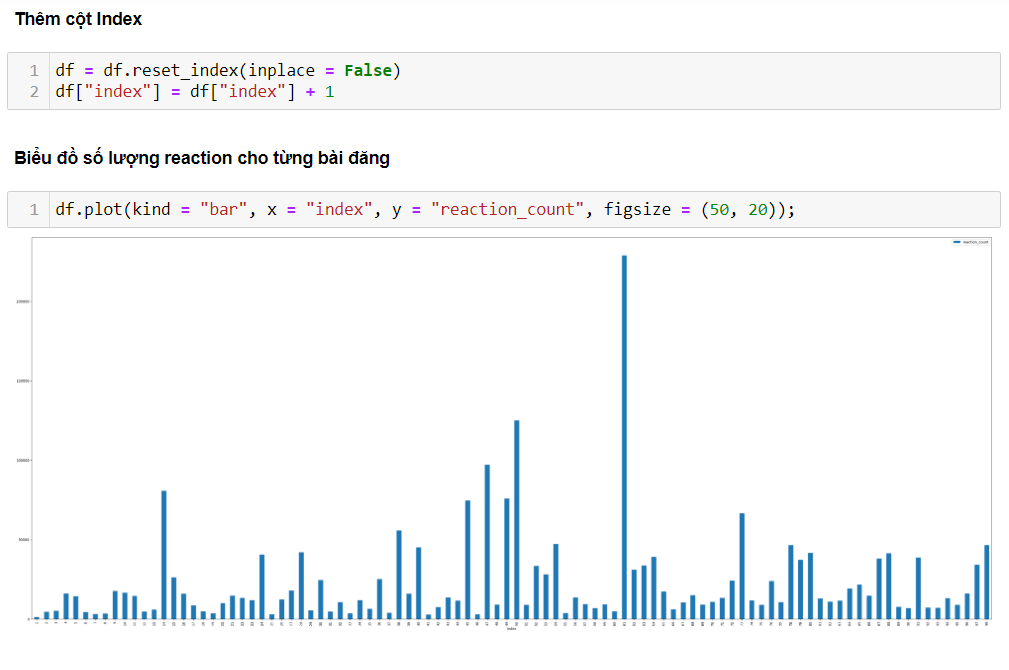
Qua biểu đồ ta thấy số comment của các bài không đồng đều trong tập dữ liệu.

3.3 Phân tích biểu cảm:

Code phần này nằm trong tệp analysis\_reactions.ipynb

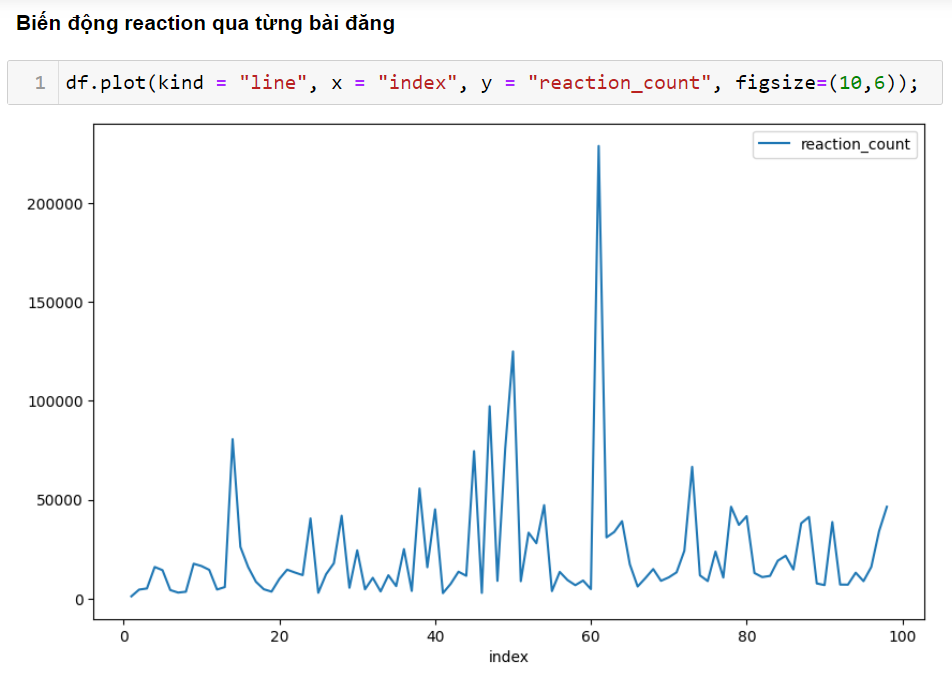
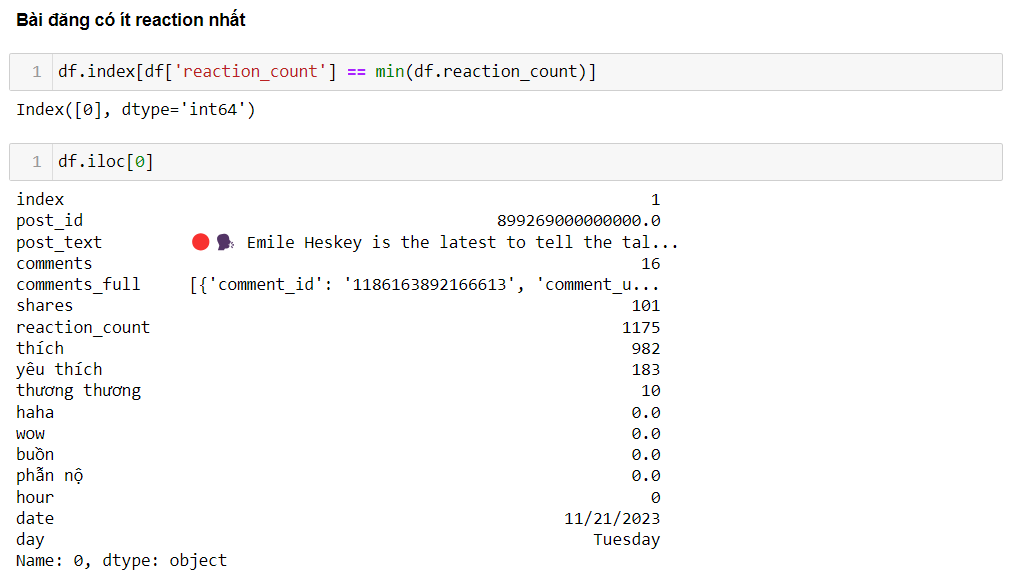
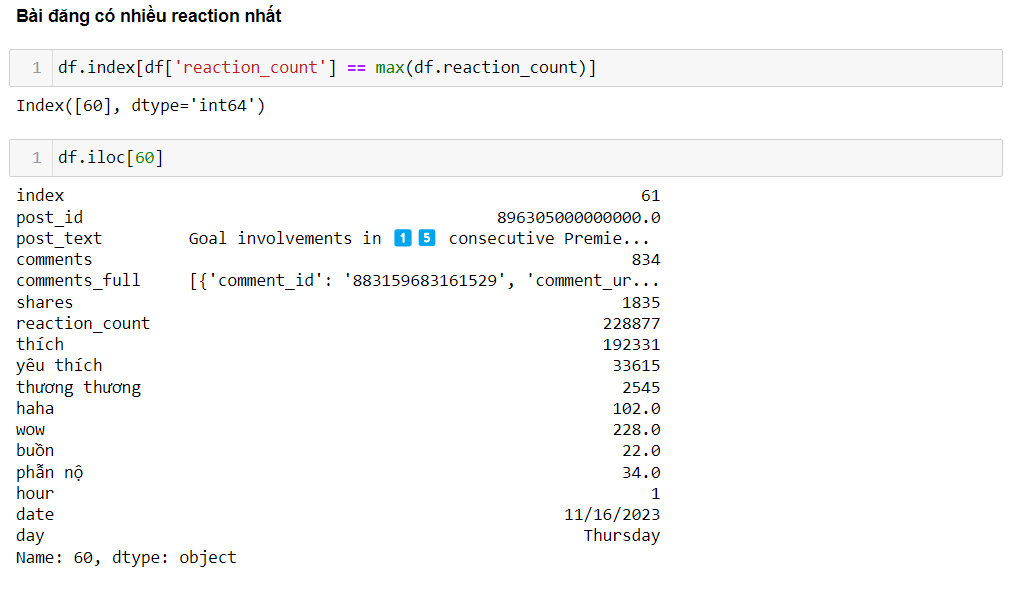
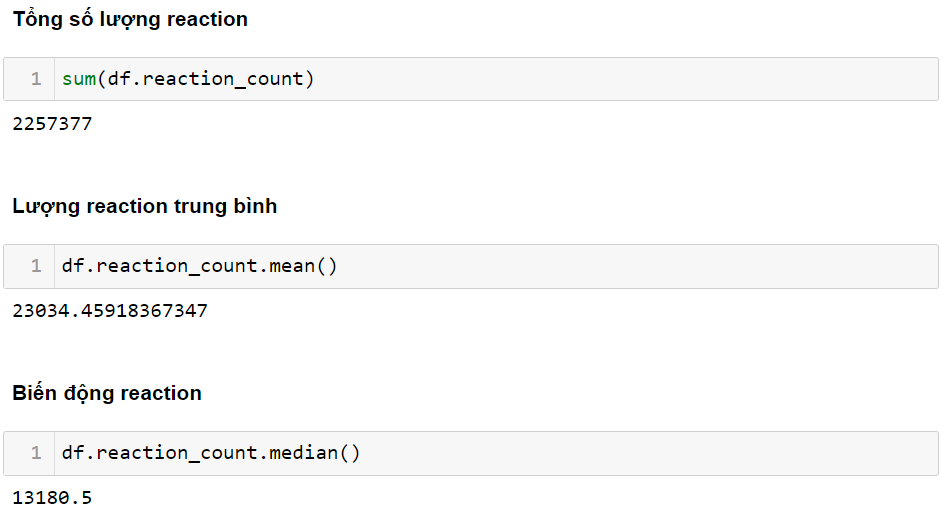


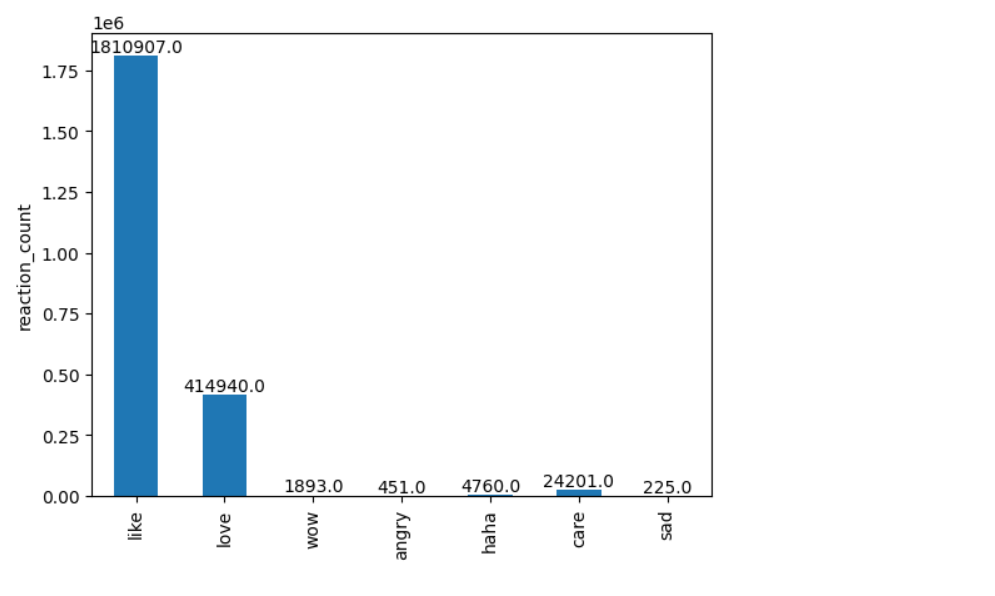
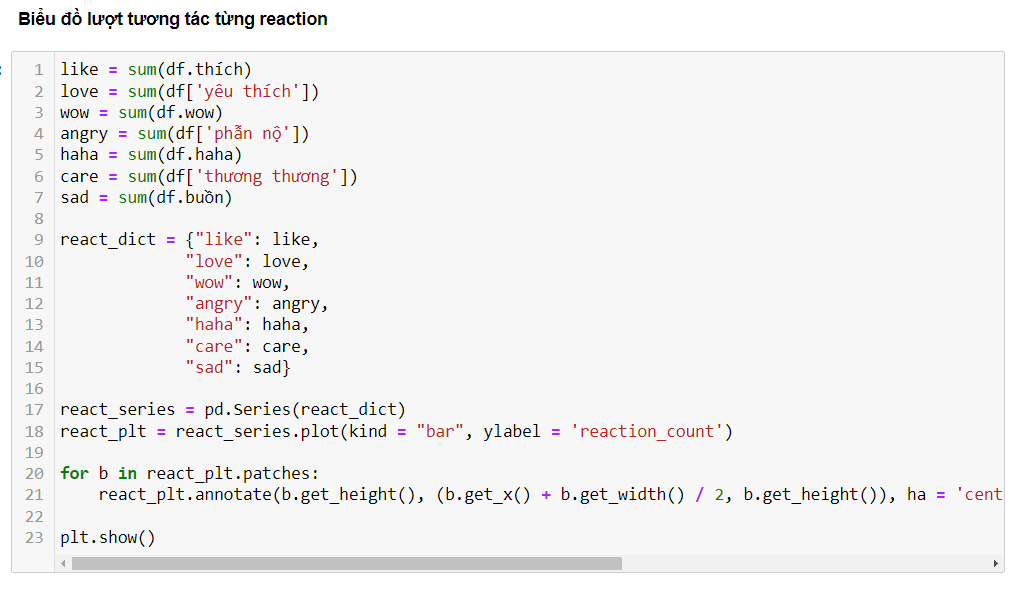




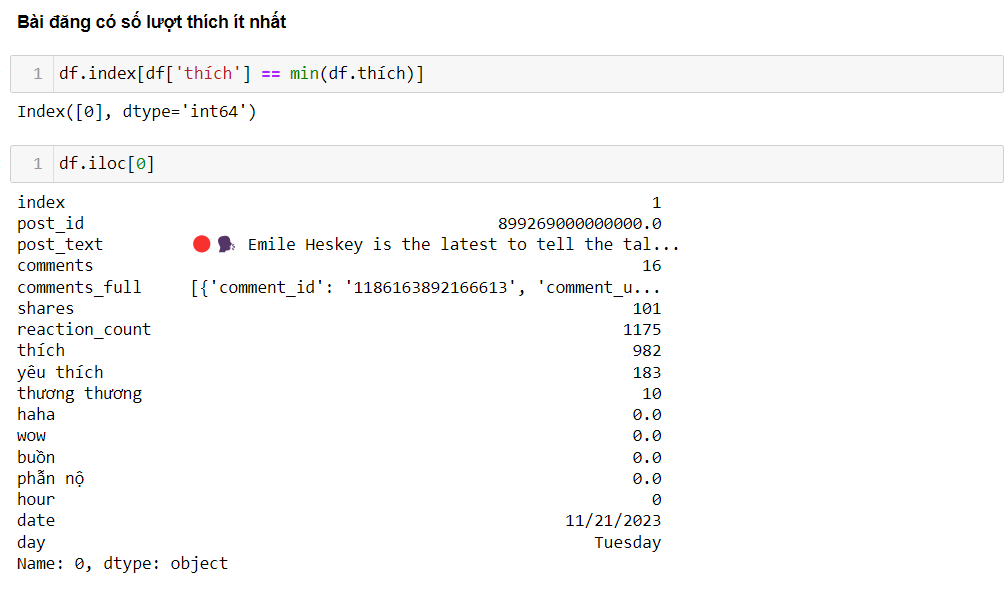
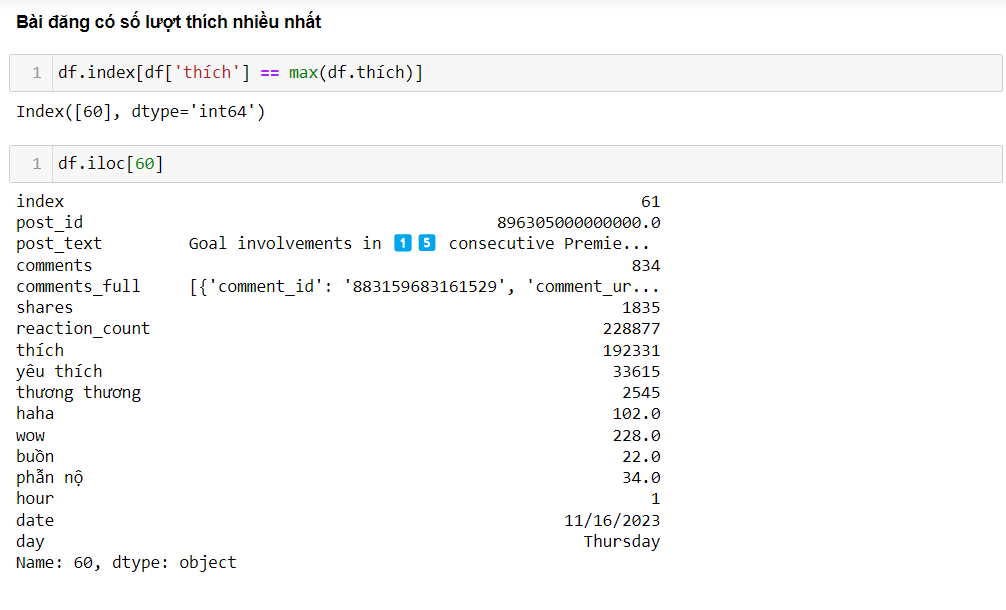
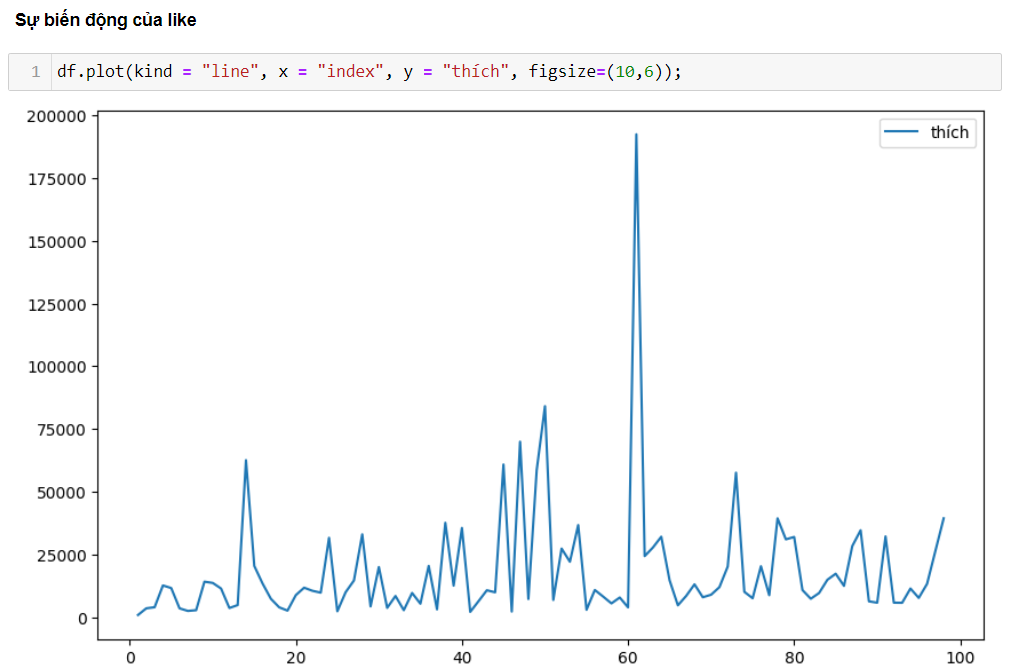
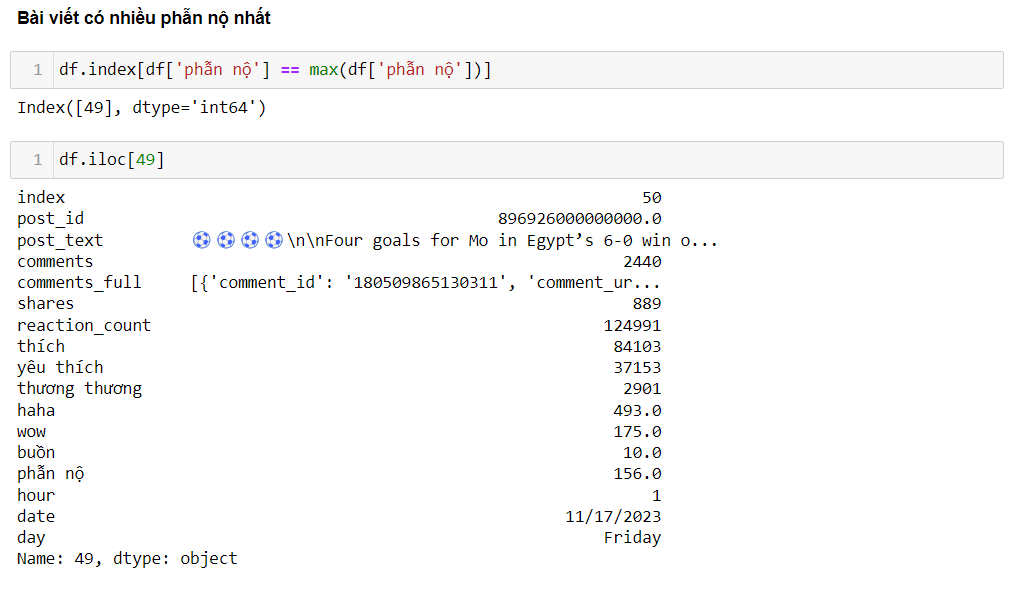
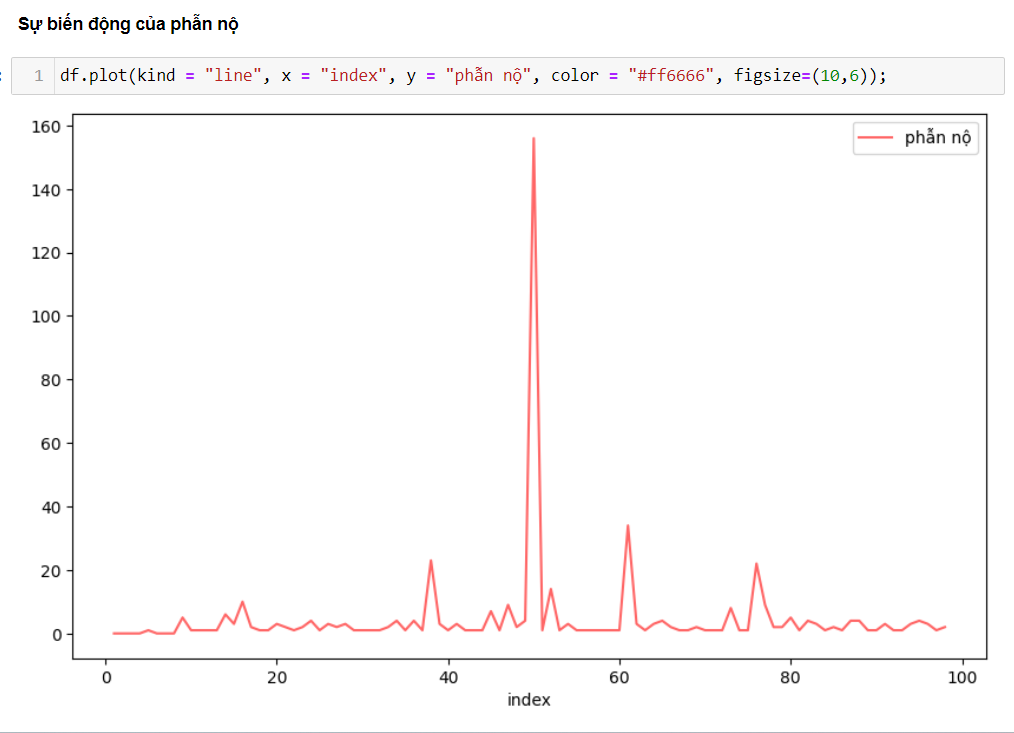
Qua biểu đồ ta thấy bài viết thứ 61 là bài viết có số lượt reaction nhiều nhất

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

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

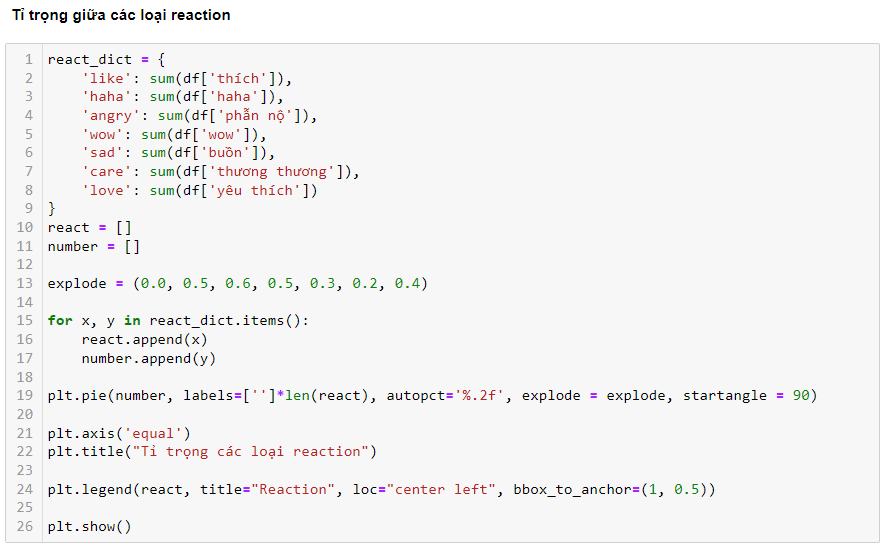


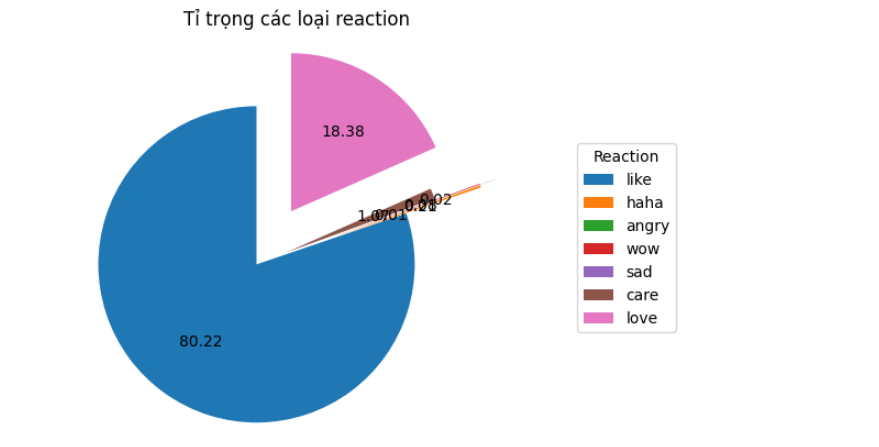
Qua biều đồ ta thấy đa số người xem chọn like và số lượt lke vượt trội hẳn so với các trường còn lại số lượt thích.

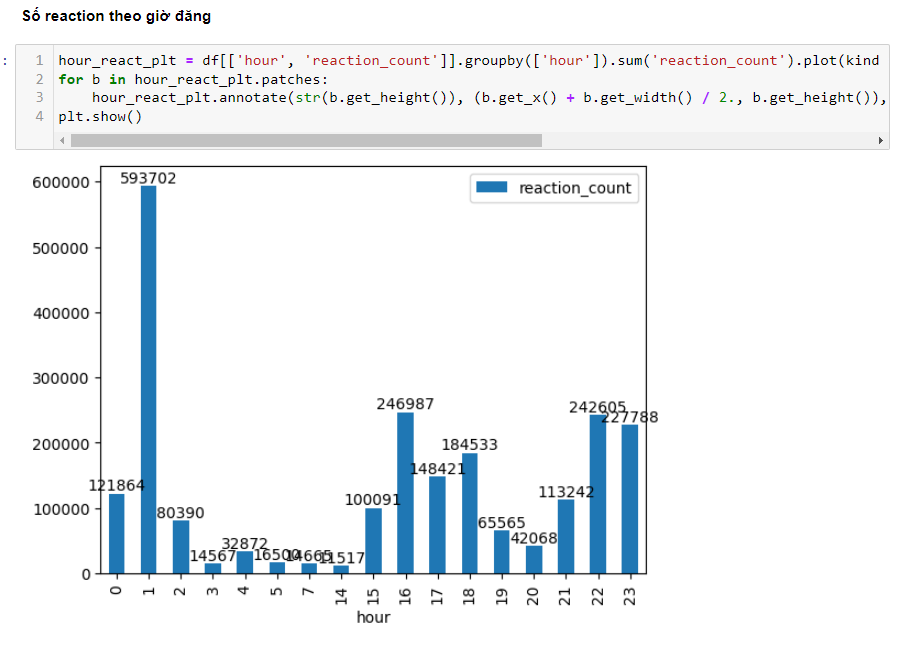
Qua biều đồ phân tích sự biến động về số phẫn nộ và số phẫn nộ lớn nhất ta thấy khi Liverpool thắng Manchester United thì fan MU ngay lập tức vào fanpage của Liverpool thả lượt phẫn nộ cho bài đăng ăn mừng chiến thắng của Liverpool.

MU quá tệ!!!!!!!!!!





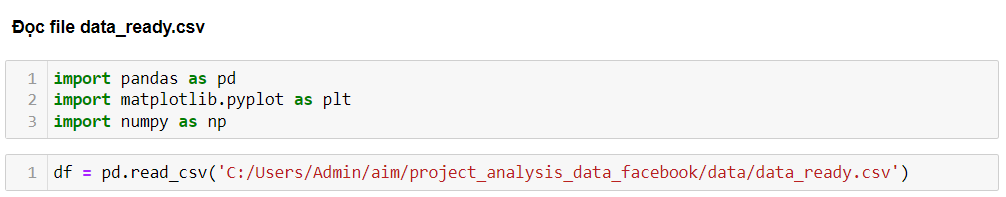
Qua biểu đồ trên ta thấy số biểu cảm haha, wow, buồn, phẫn nộ, thương thương chiếm rất ít trong tỉ trọng số biểu cảm của tất cả bài đăng.

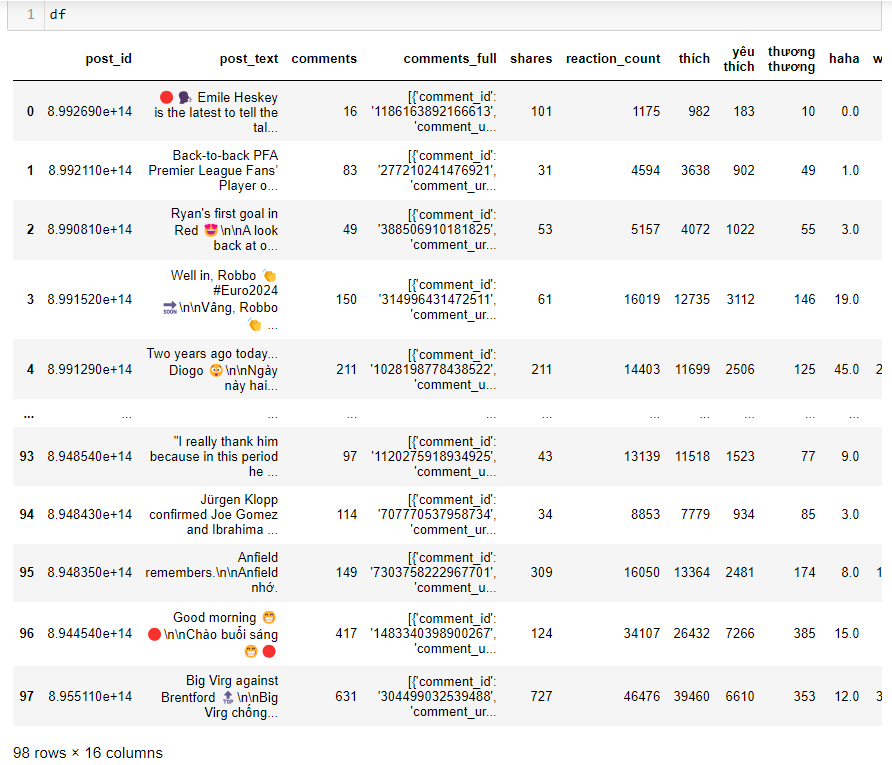


Do đội bóng nằm ở nước Anh nên đa số người theo dõi nằm ở Châu Âu, vì thế cái khung 1h theo giờ Việt Nam rất được nhiều người quan tâm nên 1h là giờ mà số lượt reaction nhiều nhất.

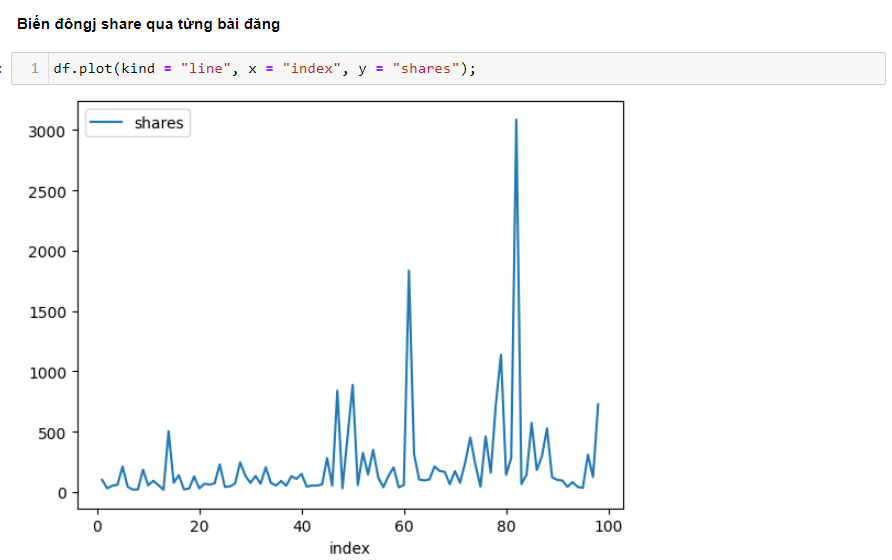
3.4 Phân tích số lươt chia sẻ:

Code phần này nằm trong thư mục analysis\_shares.ipynb





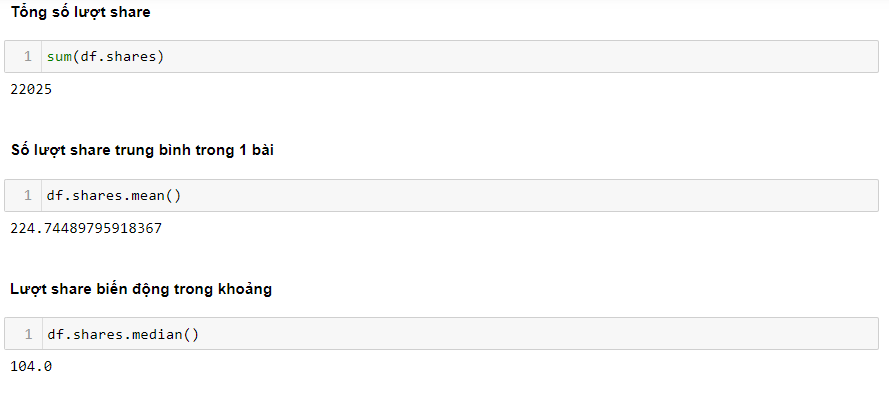




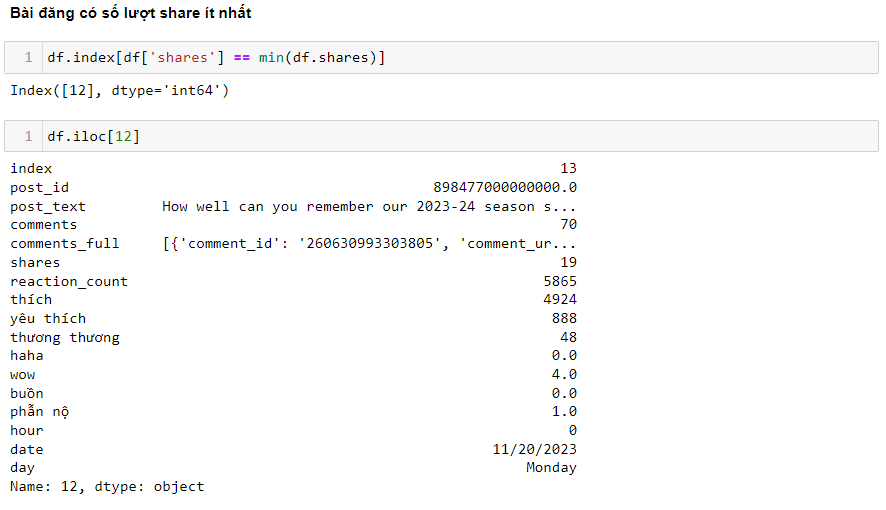
Qua biểu dồ đường trên ta thấy số lượt shares không đồng đều và đang có xu hướng tăng lên.

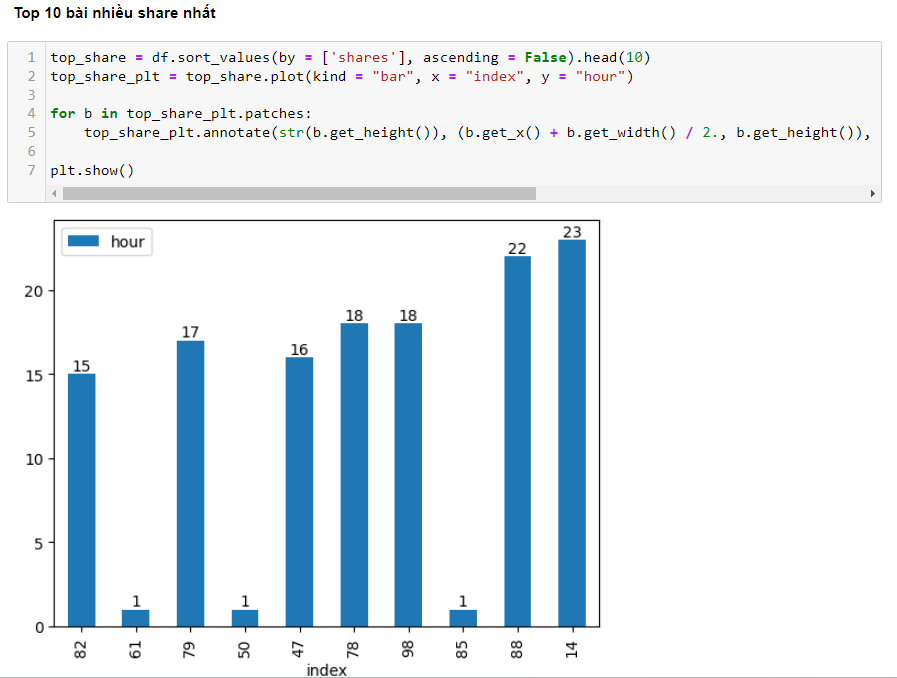


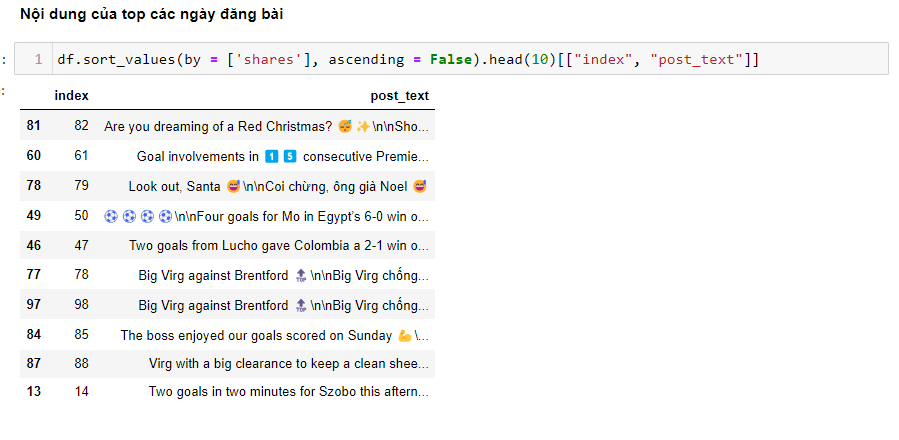
Qua biểu đồ trên ta thấy bài đăng thứ 82 có số share vượt trội.



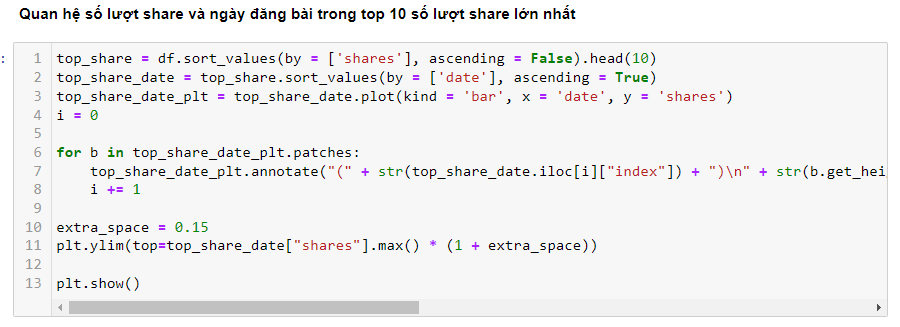


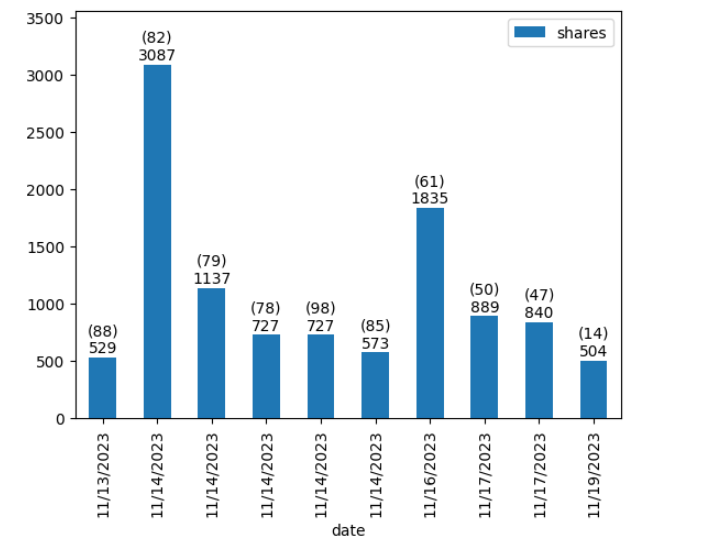




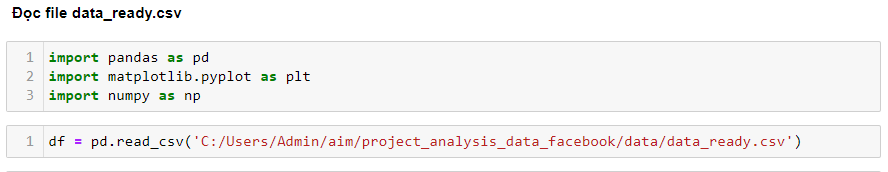


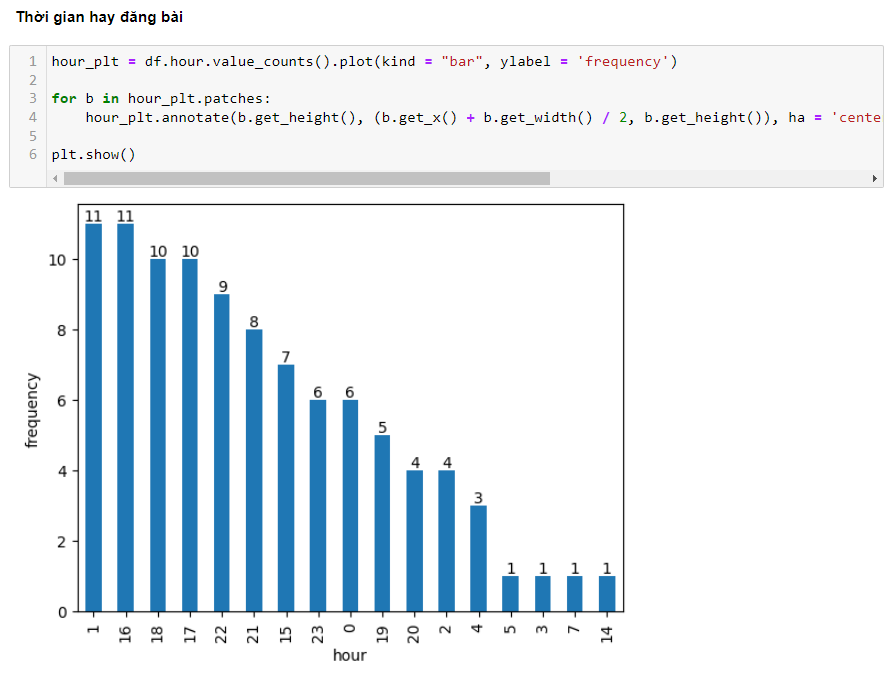




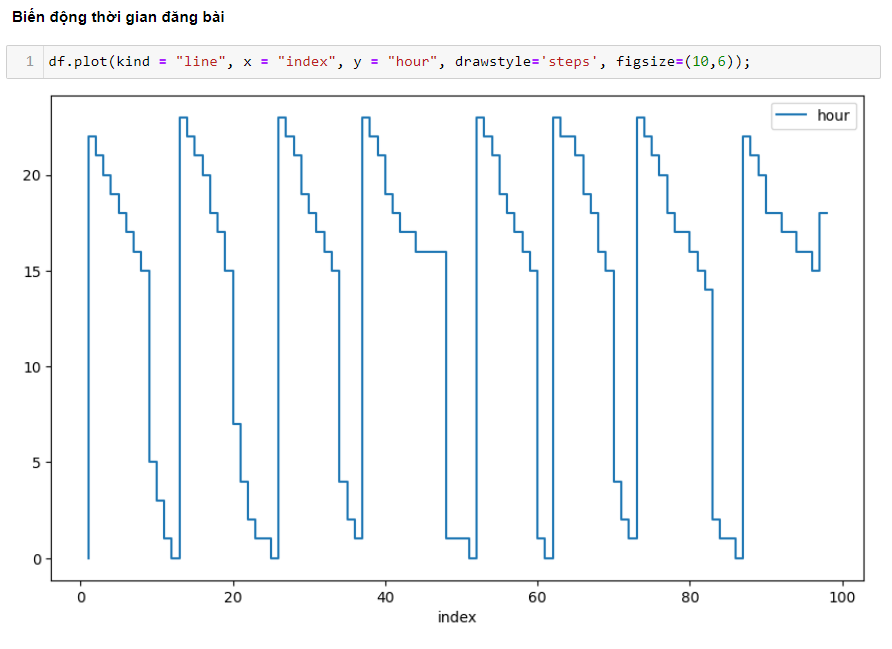


3.5 Phân tích thời gian đăng bài:

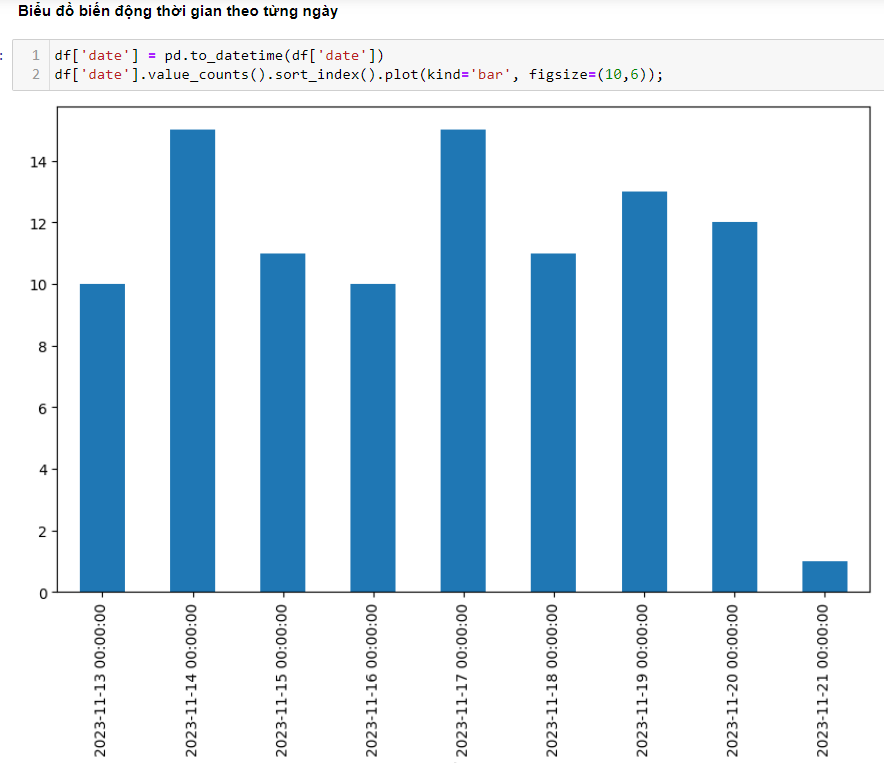




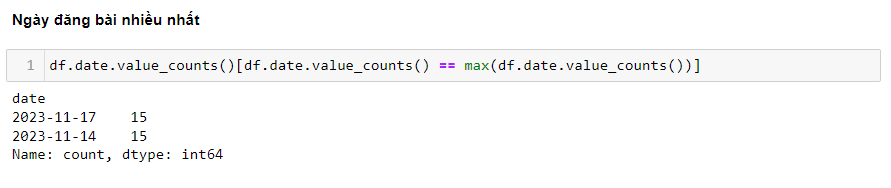
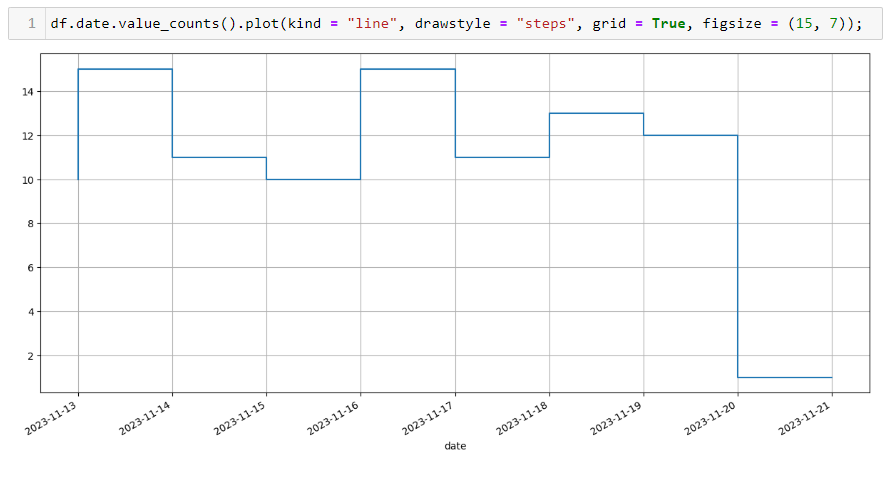
Vậy admin hay chọn các giừ như 1h, 16,17,18h để đăng bài.

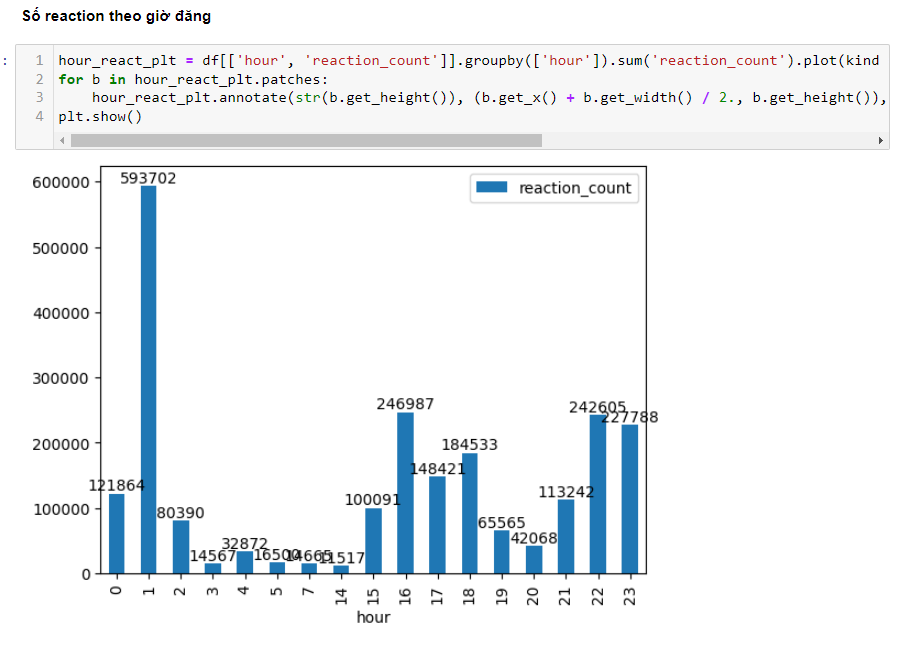


Số bài đăng theo các khung giờ chênh lệch nhau quá nhiều.



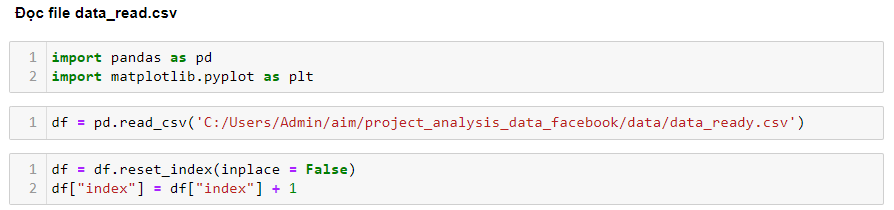
Khoảng ngày 17/11/2023 là ngày mà fanpage có nhiều bài đăng mới nhất .

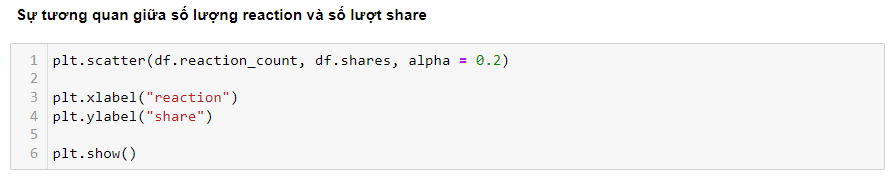


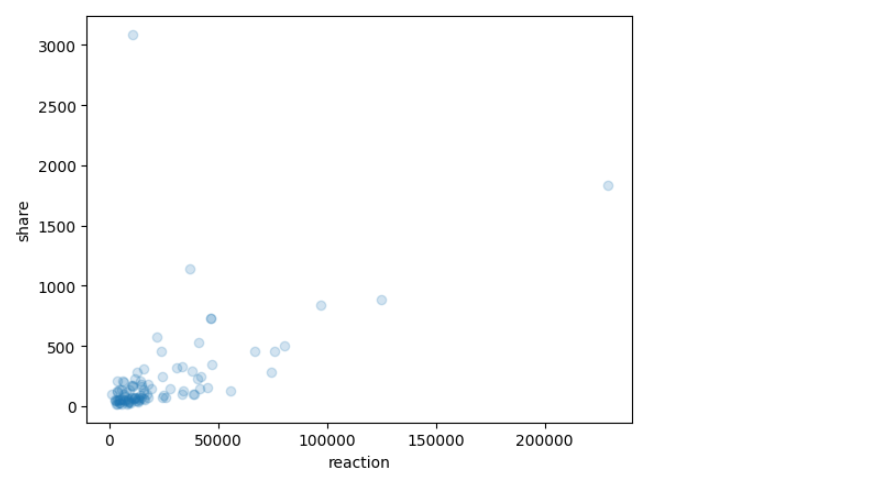


Do đội bóng nằm ở nước Anh nên đa số người theo dõi nằm ở Châu Âu, vì thế cái khung 1h theo giờ Việt Nam rất được nhiều người quan tâm nên 1h là giờ mà số lượt reaction nhiều nhất.

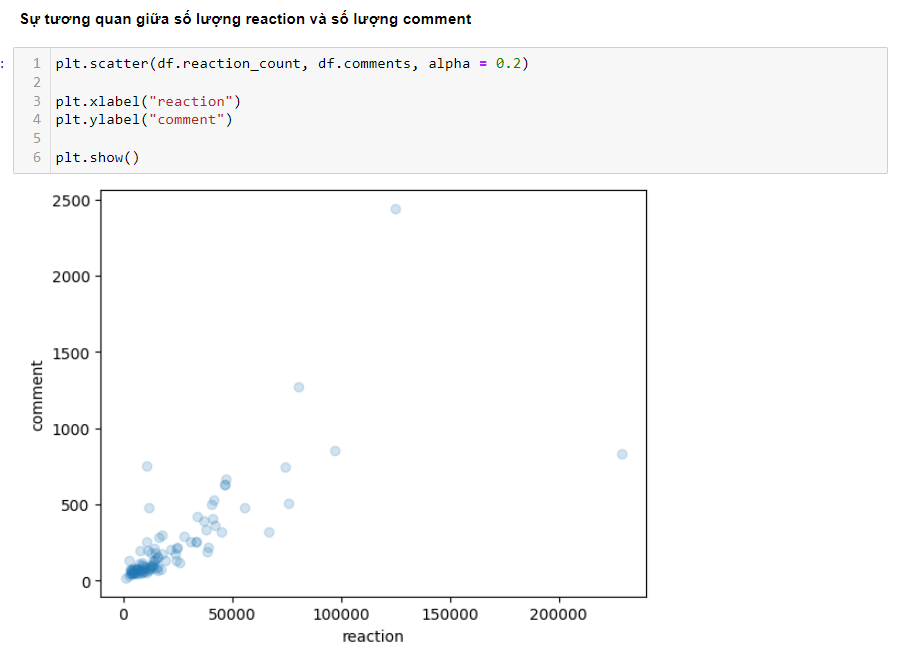
3.6 Phân tích tương quan các trường dữ liệu:



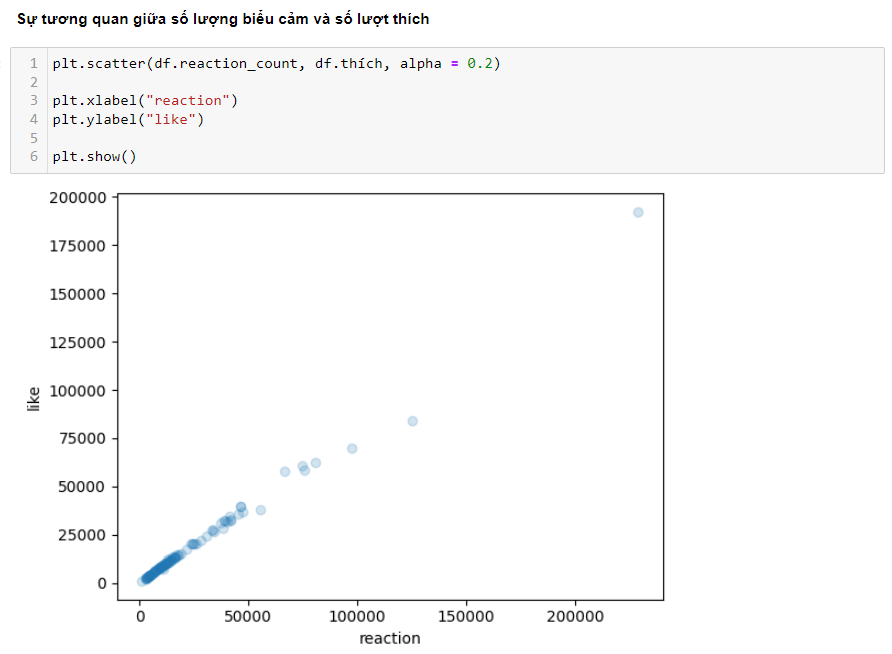




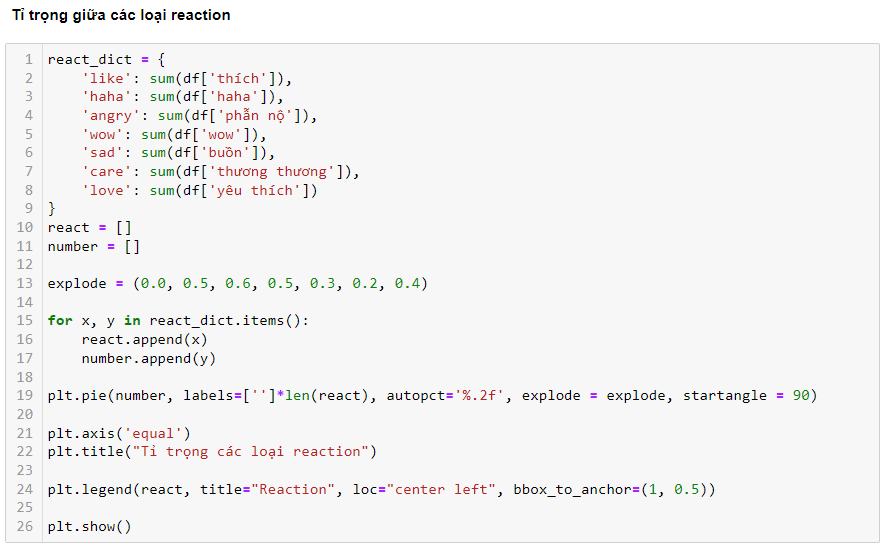
Qua biểu đồ trên ta thấy đa số các bài đăng có số lượt share không tỉ lên thuận với số lượt reaction.

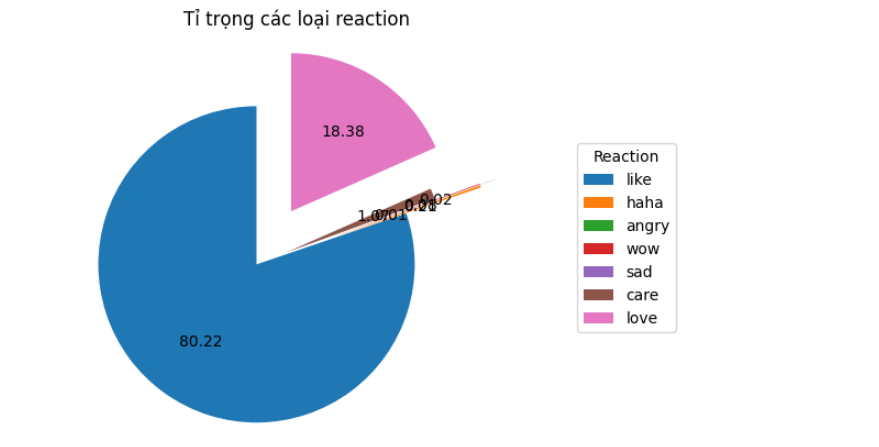


Qua biều đồ trên ta thấy số lượt bình luận tỉ lệ thuận với số lượt thả biểu cảm.

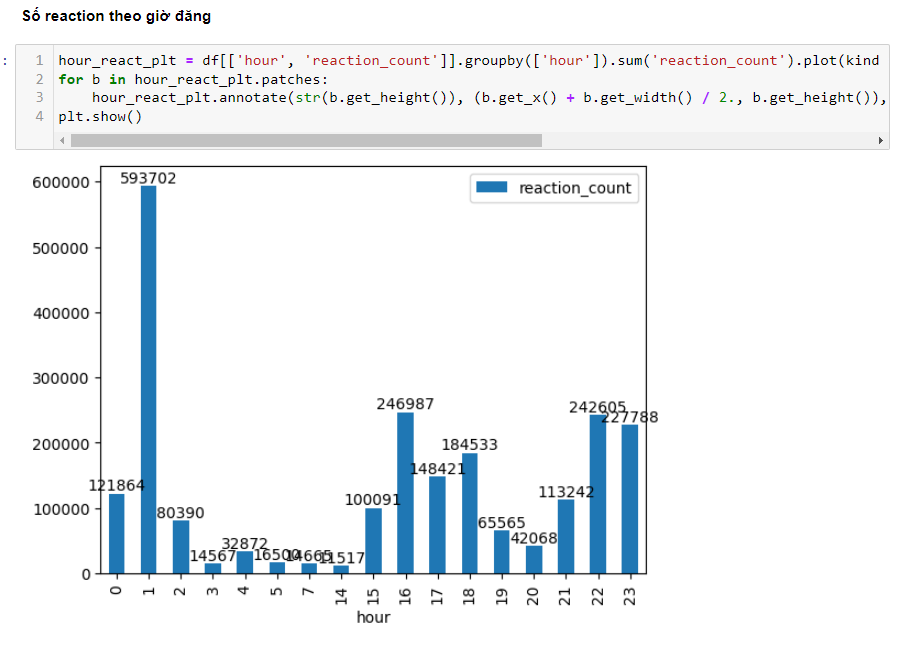


Do số lượt thích chiếm đa số trong số lượt biểu cảm nên là số lượt thích tỉ lệ thuận với số lượng reaction.

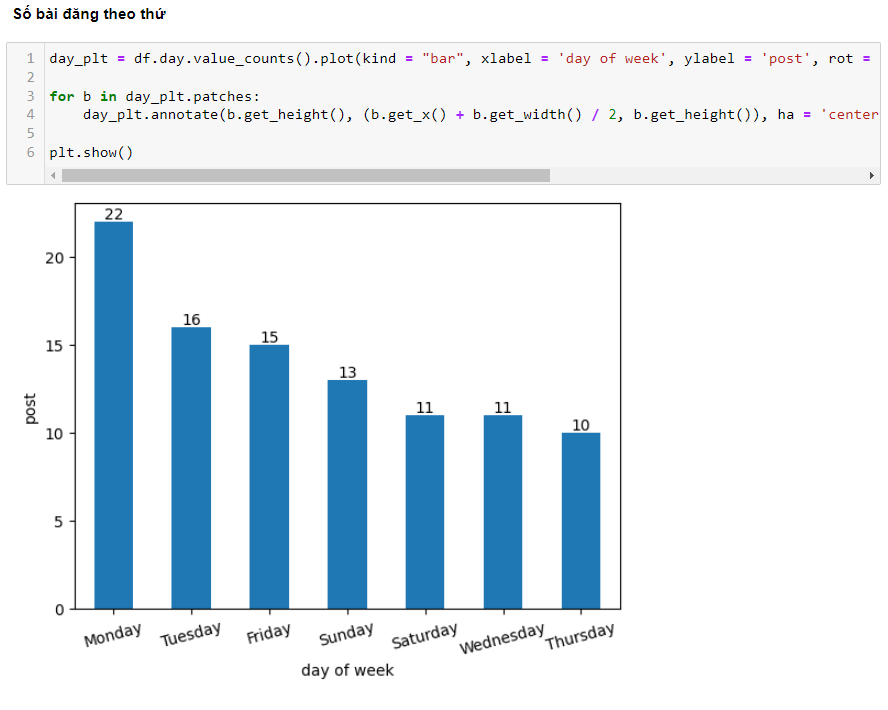




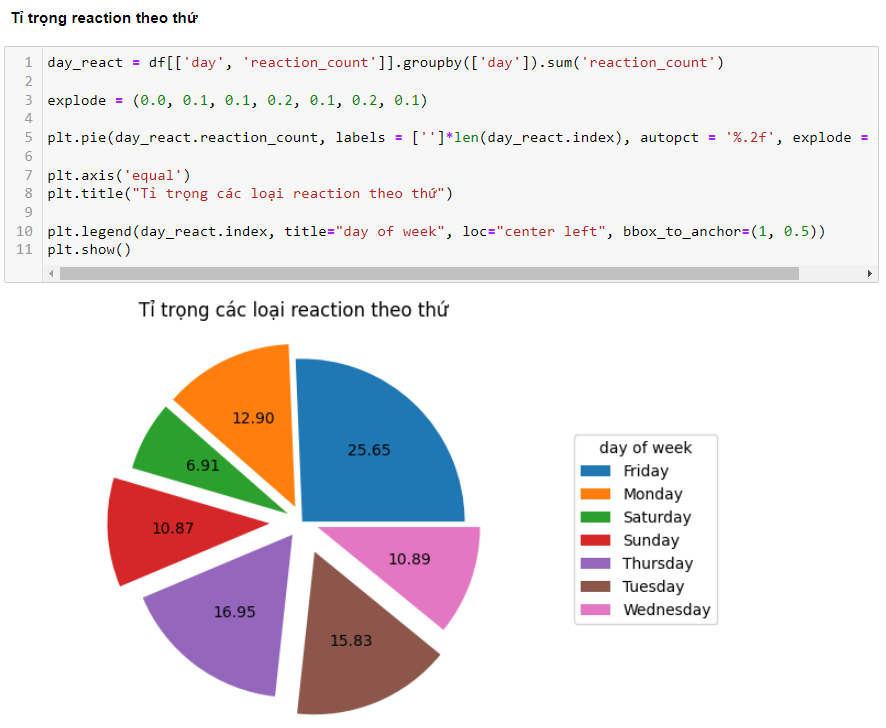
Qua biểu đồ trên ta thấy số biểu cảm haha, wow, buồn, phẫn nộ, thương thương chiếm rất ít trong tỉ trọng số biểu cảm của tất cả bài đăng.



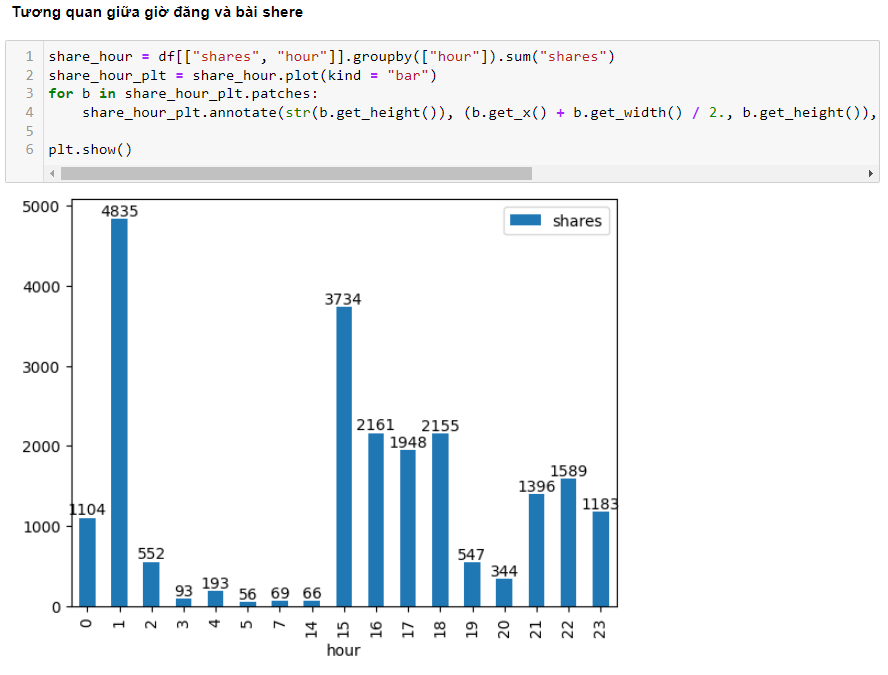
Do đội bóng nằm ở nước Anh nên đa số người theo dõi nằm ở Châu Âu, vì thế cái khung 1h theo giờ Việt Nam rất được nhiều người quan tâm nên 1h là giờ mà số lượt reaction nhiều nhất.



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

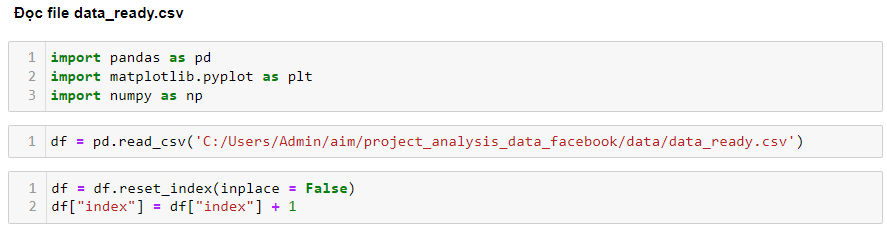


Qua biểu đồ trên ta thấy vào chủ nhật thì có rất nhiều reaction vì chủ nhật mọi người được ở nhà nên thời gian họ dành thời gian cho mạng xã hội khá cao.

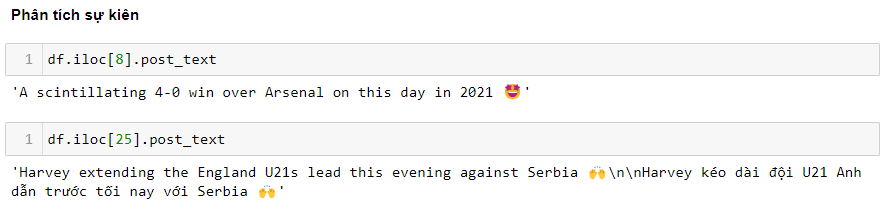


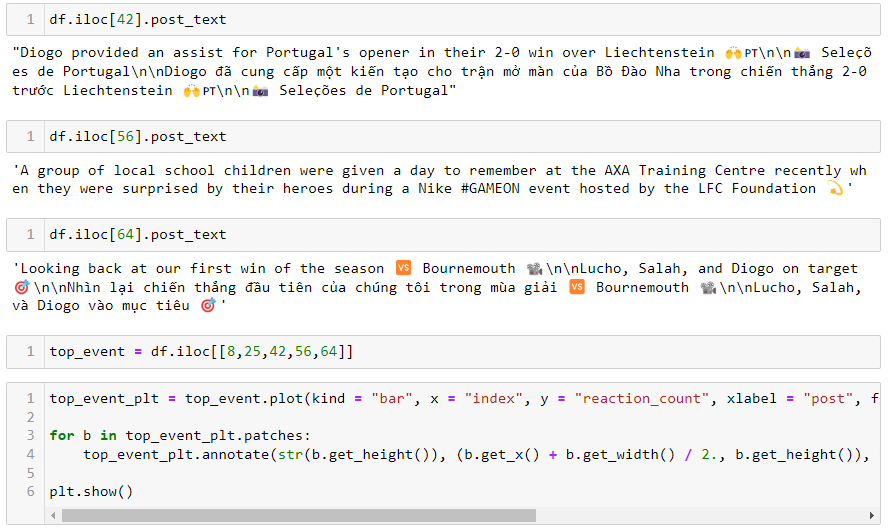
3.7 Phân tích thú vị:

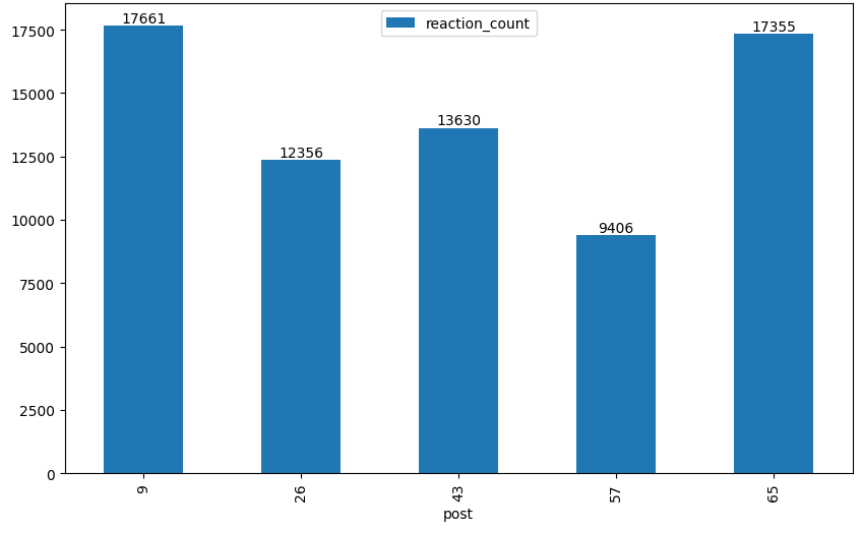
Code phần này nằm trong tệp analysis\_interesting.



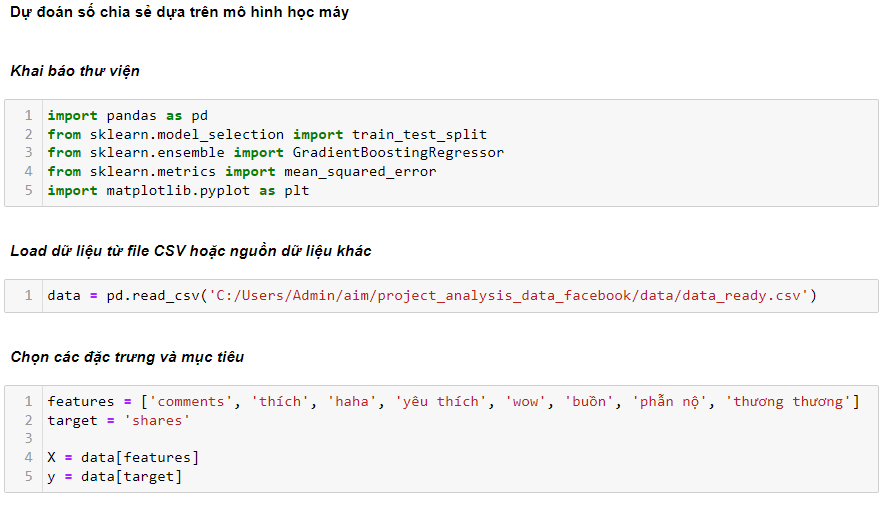


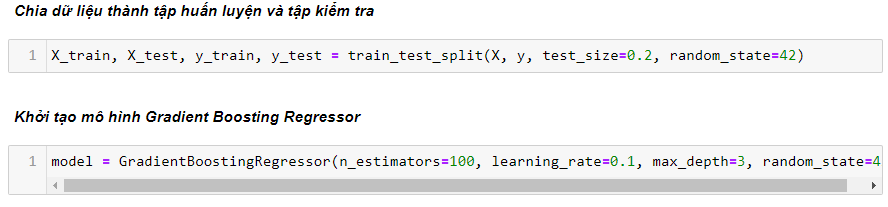




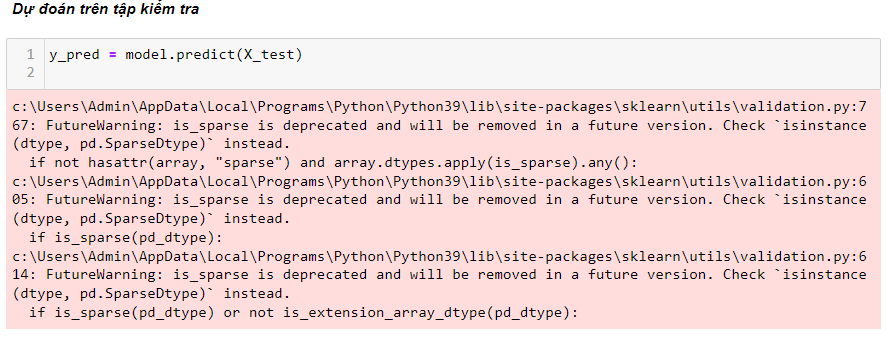


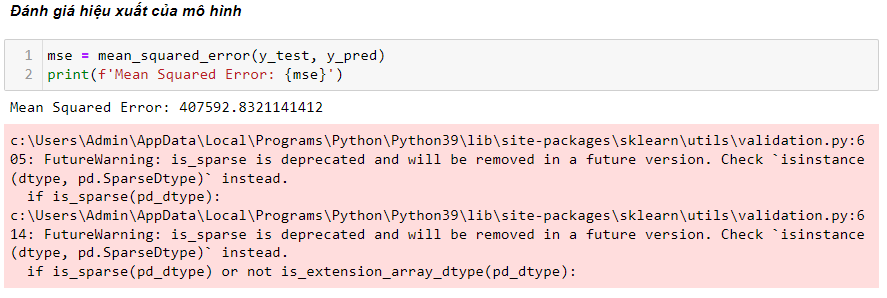
Qua biểu đồ trên ta thấy khi thấy những gì liên quan đến Liverpool thắng thì mọi người thả biểu cảm rất nhiều.

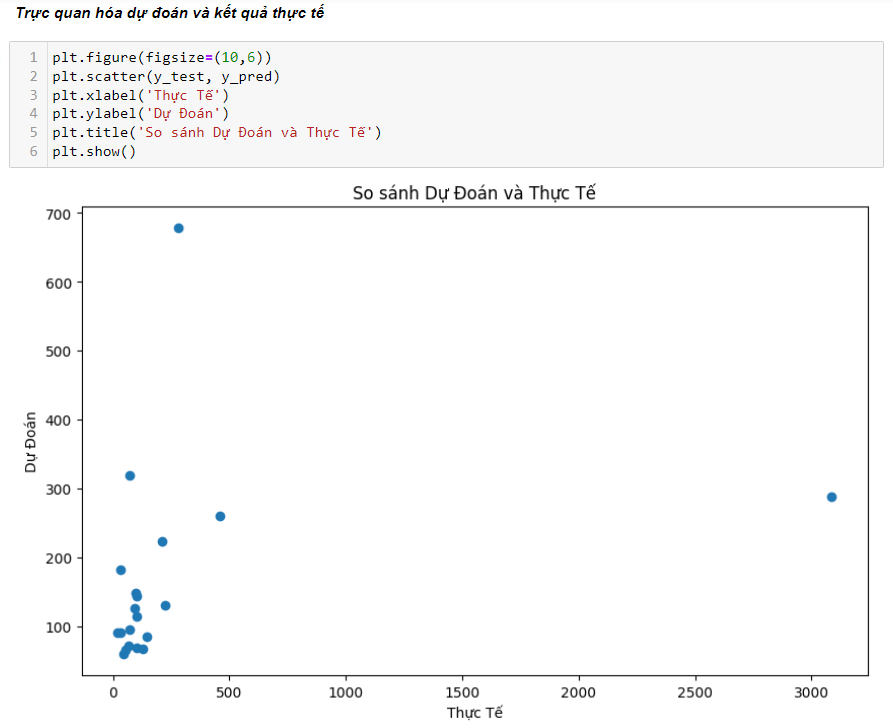








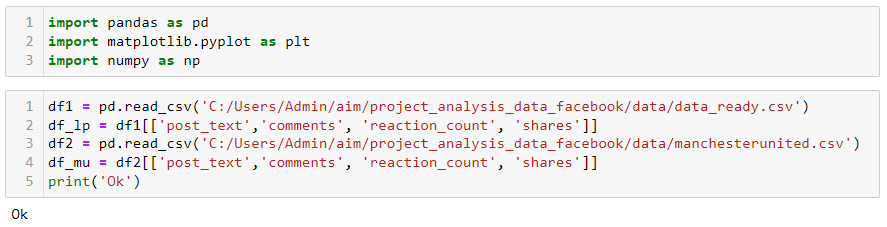


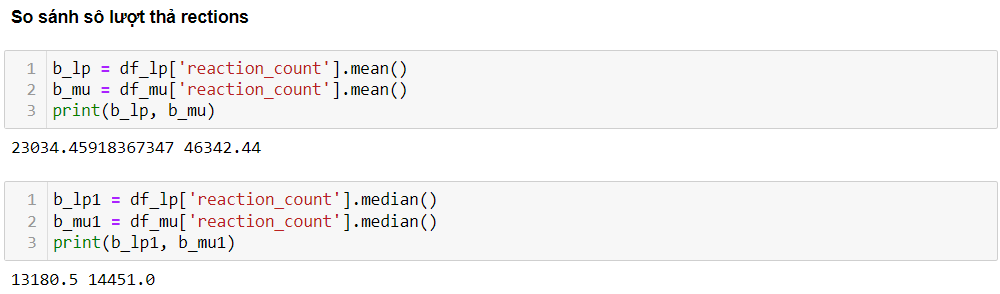


Qua biểu đồ ta thấy kết quả dự đoán số lượt share => các dữ liệu về trường share không đồng đều hoặc tập dữ liệu hơi ít.

3.8 So sánh với Fanpage Manchester United:

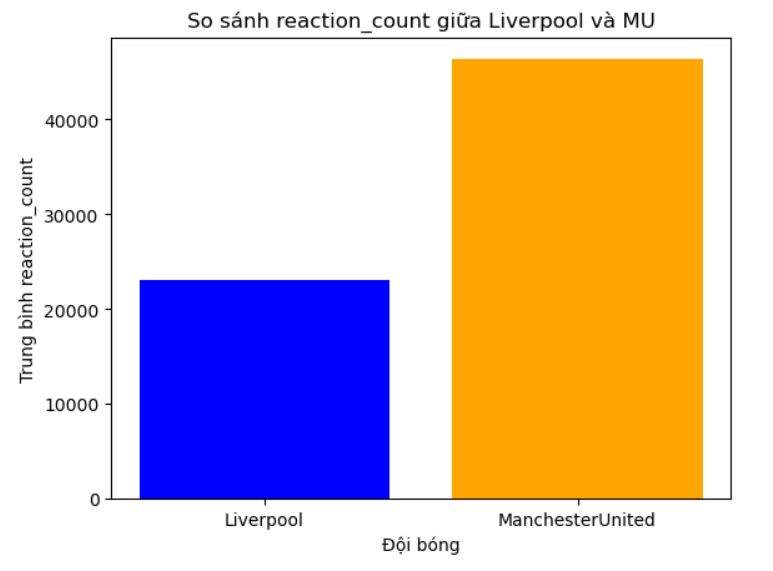
Code phần này nằm ở file analysis\_compare.csv

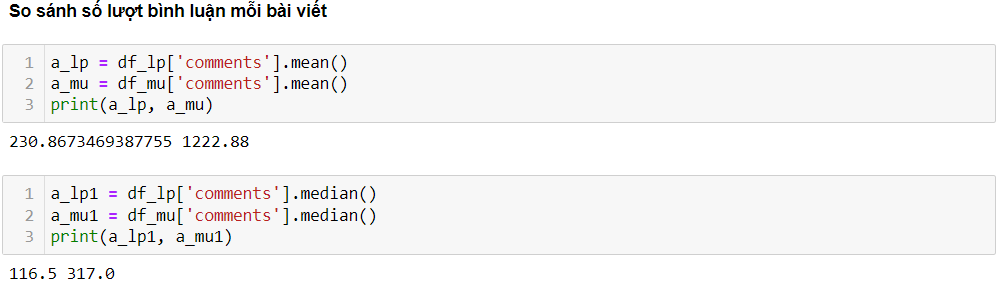




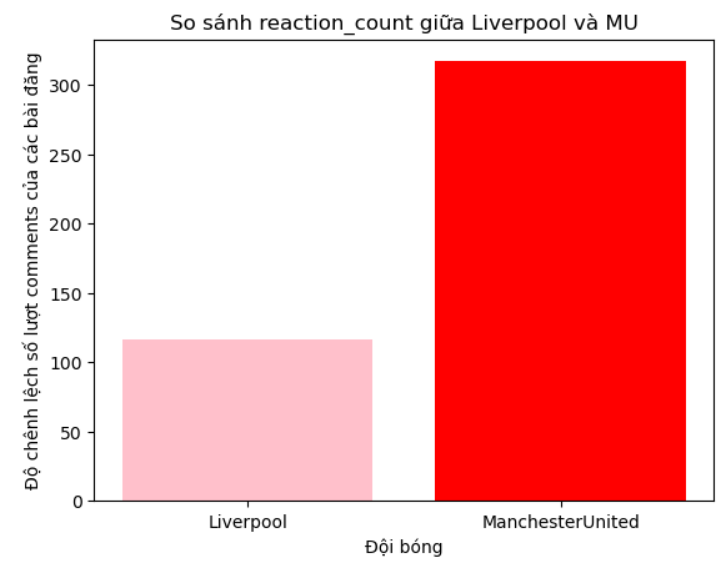
Từ kết quả trên ta thấy số reaction của MU cao hơn Liverpool rất nhiều. Cũng phải thôi fan Liverpool ít hơn hẳn mà.











Từ biểu đồ trên ta thấy độ chênh lêch của reaction qua từng bài đăng của Mu cao lơn Liverpool. Đúng mà, vì số reactions của MU cao gần gấp đôi Liverpool.





