

Task 2.2

mÔN: Đồ án kỹ thuật lập trình

GVHD: Lưu Quang Huân

Lớp L01

**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC BÁCH KHOA**

🙞🙞•⚫•🙜🙜

**Danh sách thành viên**

|  |  |  |
| --- | --- | --- |
|  |  |  |
| STT | HỌ VÀ TÊN | MSSV |
| 1 | Mai Nguyễn Minh Hiếu | 1812171 |
| 2 | Ngô Quang Anh | 1810014 |
| 3 | Hồ Ngọc Trí | 1814454 |

**1)Các công việc thực hiện**

-Mai Nguyễn Minh Hiếu :

crawl thêm dữ liệu để được ~10.000 cũng như là tập hợp dữ liệu của các thành viên trong nhóm lại cũng như là sạch dữ liệu.

-Ngô Quang Anh :

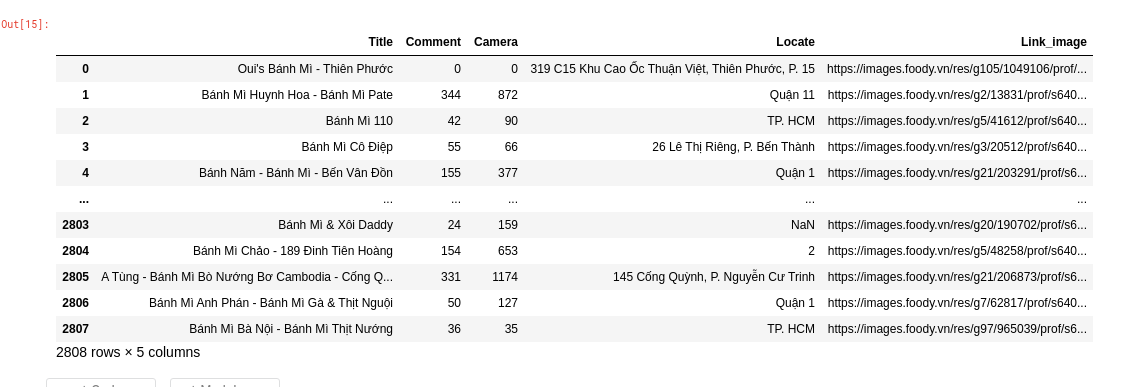
Phân tích dữ liệu và viết báo cáo

-Hồ Ngọc Trí :

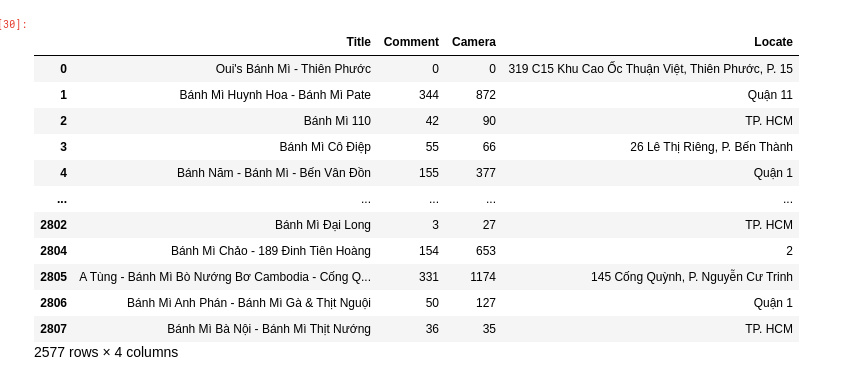
Phân tích dữ liệu và đưa dữ liệu lên Database

**2)Làm sạch dữ liệu**

Ở đây ta có thể thấy dữ liệu thô khi được crawl về từ trang web



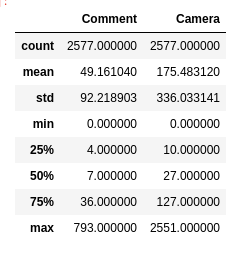
từ dữ liệu ta sẽ lược bỏ các thành phần NaN cũng như là phần link ảnh không cần thiết cho việc phân tích nên cũng lượt bỏ.



dữ liệu đã được giảm từ 2808 -> 2577

**3)Phân tích dữ liệu**

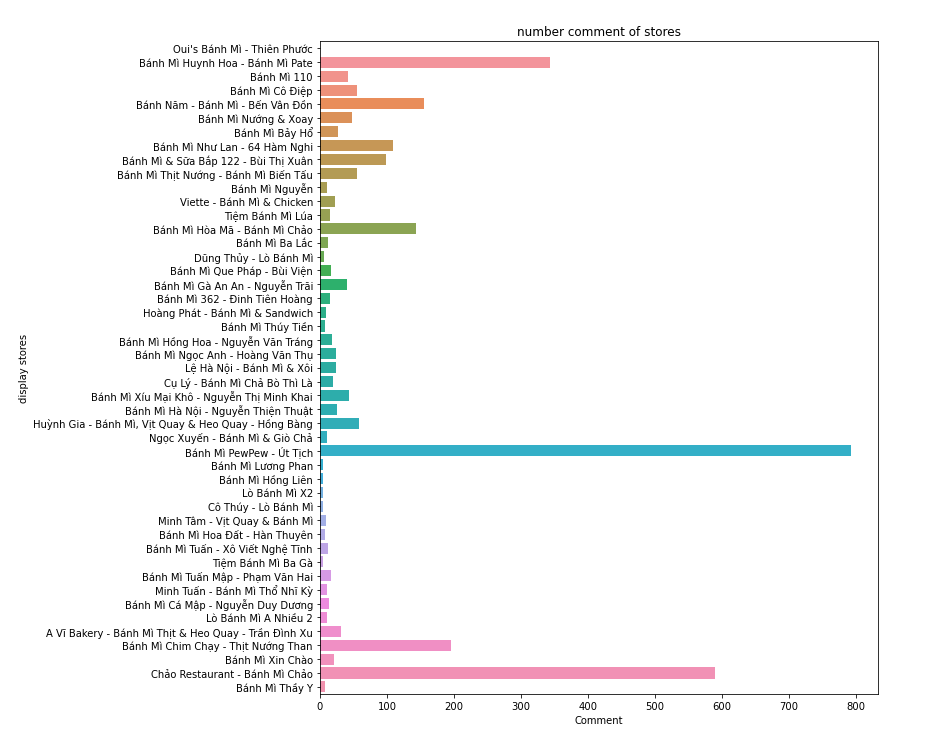
đến với dữ liệu chung



như ta có thể thấy Tổng số dữ liệu sau khi được lọc thì còn : 2577 data

tổng trung bình Comment 49 khá là cao cho mỗi cửa hàng cũng như là có 175 Photos được thêm vào cửa hàng.

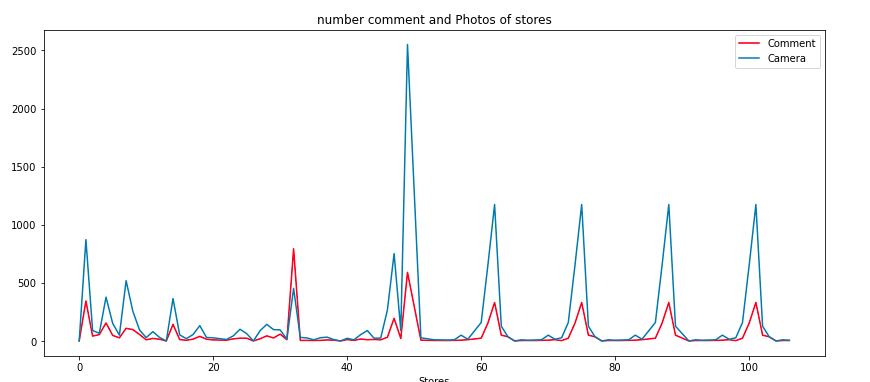
Ở đây ta có thể thấy được có cửa hàng không có Comment hay Photos nhưng cũng có cửa hàng có tận 793 Comment hay 2551 Photos dẫn đến sự chênh lệch dữ liệu quá lớn.



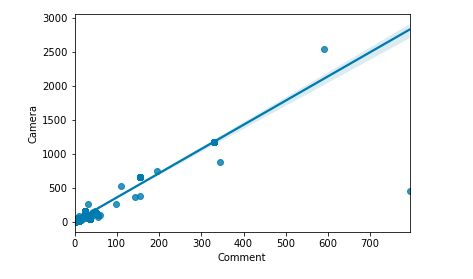
đây là các 50 cửa hàng bánh mì được hệ thống chọn đầu tiên.Theo biểu đồ ta thấy 'Comment' cho các cửa hàng là không đồng đều,việc không đồng đều này có thể là do vị trí địa lý , hay cũng có thể là do độ nổi tiếng của các cửa hàng.Đặc biệt trong đó có cửa hàng

"Bánh Mì PewPew - Út Tịch" là chiếm lượng 'Comment' nổi trội hơn nhất (~800).

Tiếp theo ta sẽ cùng xem lượng ảnh của cửa hàng được chụp cũng như lượng Comment:



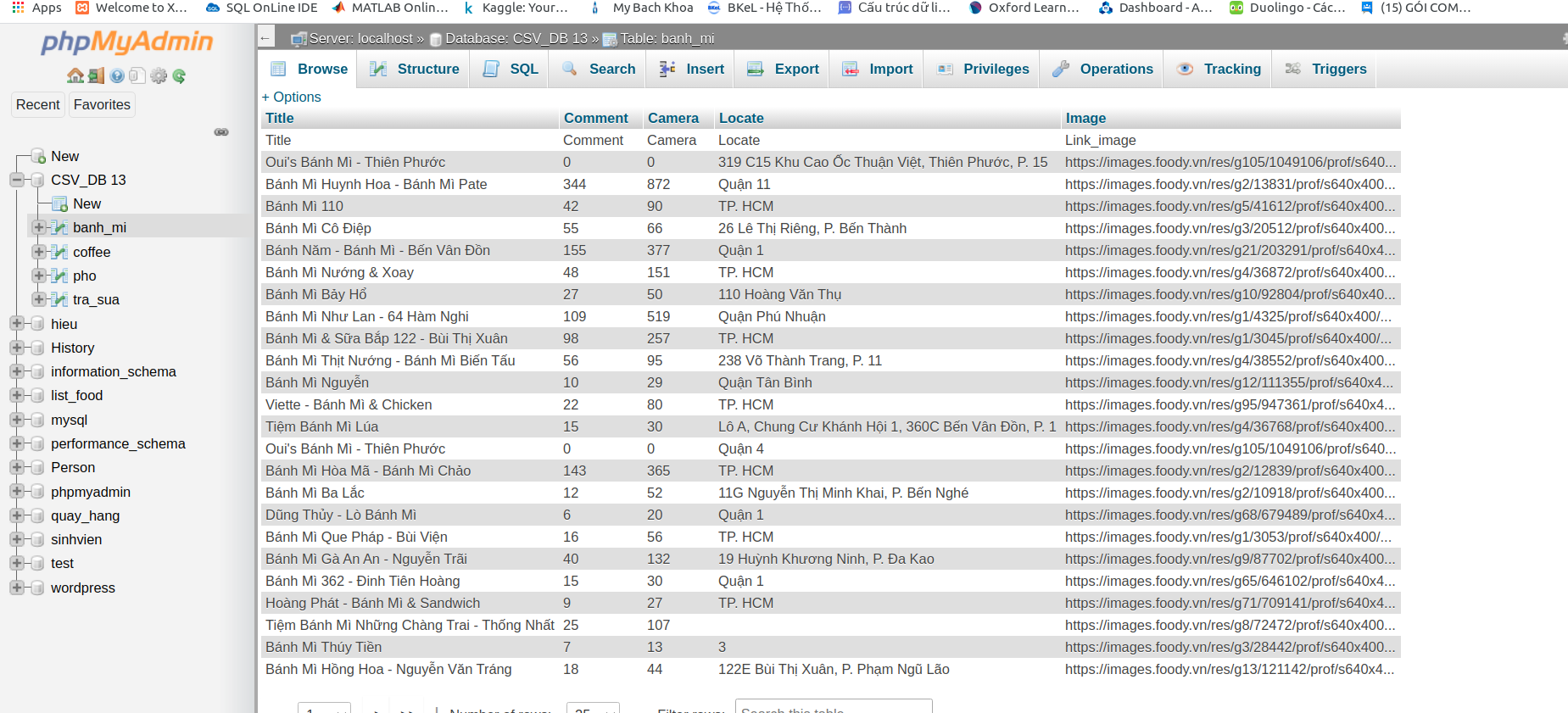
ta thấy lượng "Comment" ít hơn hẳn so với các lượng Photo đã được chụp.Chứng tỏ khách hàng thích review bằng ảnh hơn so với Comment cũng như là thể hiện được chân thật phong cảnh của cửa hàng.Lượng ảnh ở những cửa hàng nhất định luôn có những đột biến có khi lên đến (>2500 Photos)



ở đây là lượng phân bố Comment và các Photos ta có thể thấy có chiều hướng tăng dần khi mà các lượng Comment càng nhiều thì lượng Photos của 1 cửa hàng cũng được tăng đáng kể (không kể đến các giá trị quá lớn vì nó có thể gây nhiễu ) nhưng nhìn chung thì ta có thể thấy lượng Comment cũng có ảnh hưởng đến cửa hàng

**4)Đưa dữ liệu lên trên Database**

Ngoài ra nhóm cũng đã thành công trong việc đưa các dữ liệu lên trên Database với tổng ~10.000 data để chuẩn bị cho việc ứng dụng sau này.



**5) Dự kiến công việc tuần tới**

-Crawl thêm các dữ liệu cần thiết cho việc ứng dụng hơn

-Tìm hiểu ứng dụng của data crawl được

-ứng dụng cơ bản với các data crawl được trên Database