VIET NAM NATIONAL UNIVERSITY HO CHI MINH CITY

TRUONG DAI HOC KHOA HOC TU NHIEN
KHTN
TP. HO CHI MINH

Faculty of Information Technology

Môn học : Khoa Học Dữ Liệu
Lớp CQ2016/2 – Học kỳ I/2019-2020
GV: Trần Trung Kiên

BÁO CÁO LẦN I- ĐỒ ÁN CUỐI KỲ

<Tiến độ thực hiện đồ án tính đến ngày 02-12-2019>

Nhóm 14: *Đặng Phương Nam – Lê Minh Nghĩa*

NỘI DUNG CHÍNH

Câu hỏi đặt ra?

Thu thập dữ liệu

Cho các thông tin về một căn nhà
(đường, huyện/quận, mặt tiền, đường vào, diện tích, nội thất, số tầng,…)

Căn nhà này có giá là bao nhiêu tiền?

CÂU HỎI ĐẶT RA?

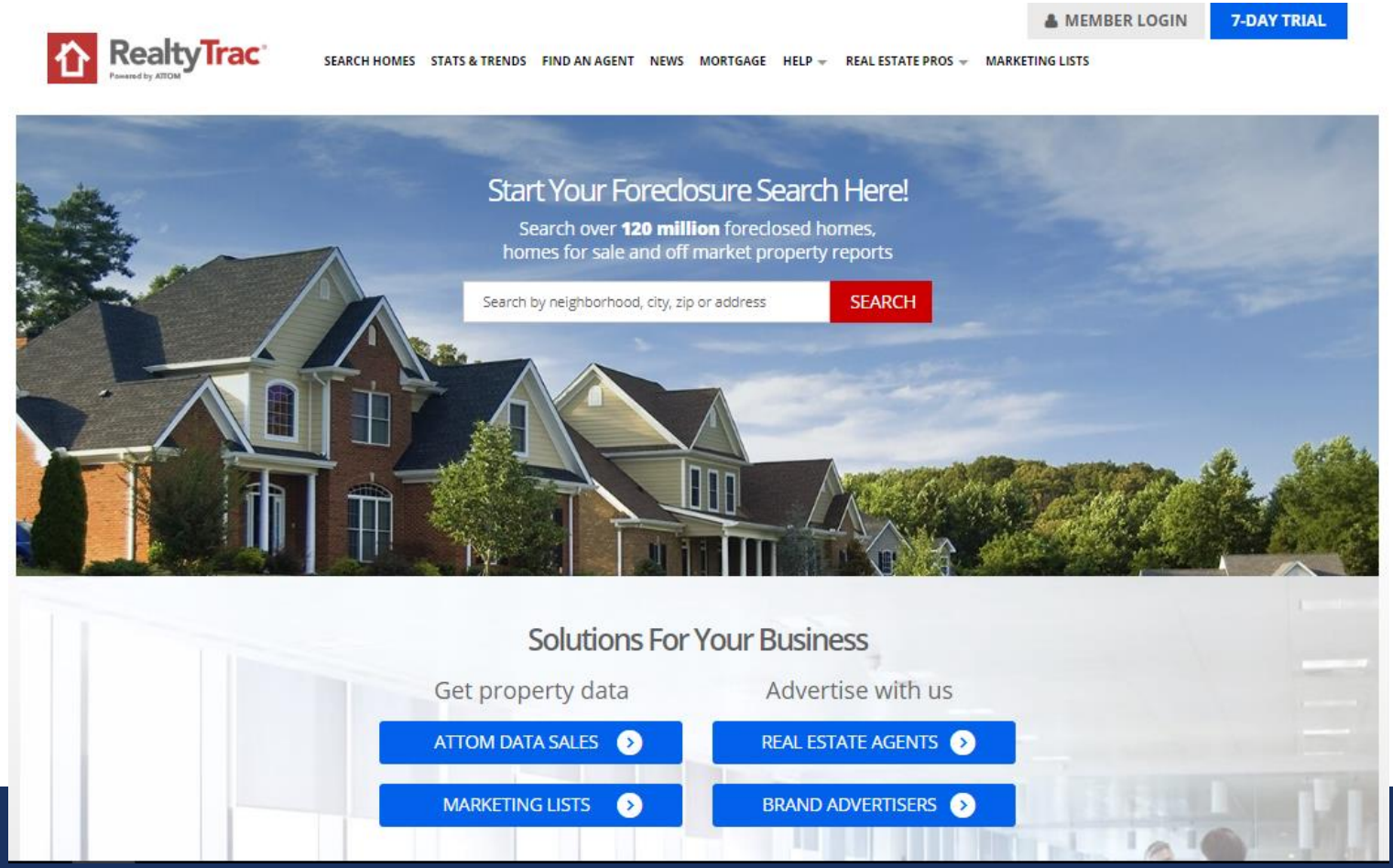
Người bán có thể dự đoán được giá trị căn nhà mà mình muốn bán.

Người mua có thể ước lượng được căn nhà mình muốn mua có giá cả hợp lý hay không?

Nguồn gốc câu hỏi: nhóm tự nghĩ ra.

LỢI ÍCH CỦA CÂU HỎI

Trang web thu thập dữ liệu: https://www.realtytrac.com/

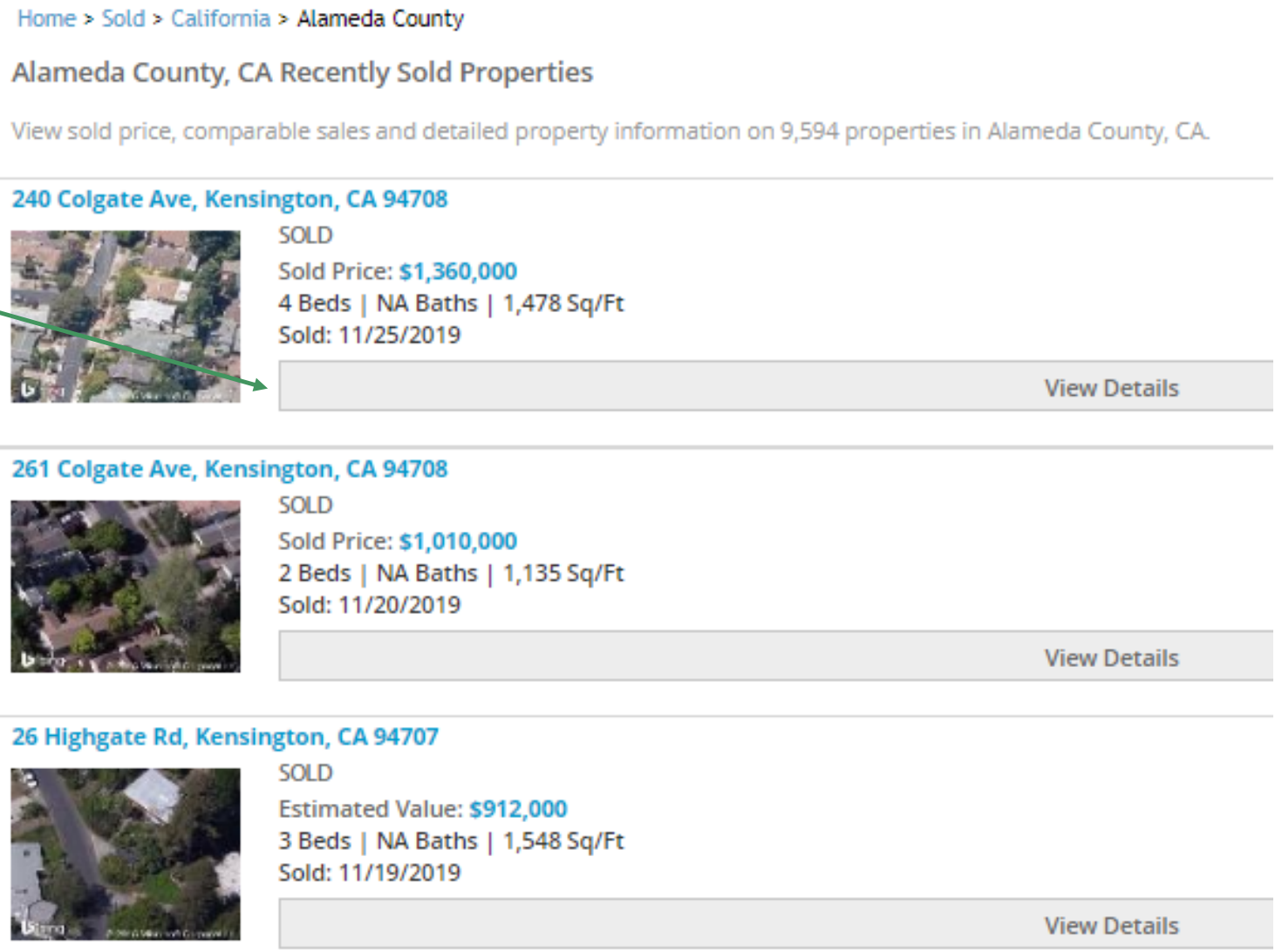


THU THẬP DỮ LIỆU – PARSE HTML

Thu thập dữ liệu về danh sách "Các căn nhà đã được bán thành công tại bang California của nước Mỹ"

Parse HTML để lấy:

+ Detail url.



Ở mỗi lần parse tại bản tin rao bán của trang web nhóm đều thực kiểm tra việc lấy dữ liệu có hợp pháp hay không tại file robots.txt của trang web:

https://m.realtytrac.com/robots.txt

Ở mỗi bản tin bán nhà thành công, nhóm muốn lấy thêm một số thông tin chi tiết:

Parse HTML để lấy:

+ Date Sold.
+ Mortgage.
+ Address.

+ Property Infor:
  - Bedroom.
  - Bathroom.
  - Size.
  - ….

Property Info

Type	Single Family Residence
Bedrooms	4
Bathrooms	2
Size	1,478 sqft
Lot Size	5,000 sqft
Year Built	1938
Sold Price	$1,360,000
Property ID	1102358090
County	Contra Costa County, CA
Parcel Number	5701800079

Ở mỗi lần parse tại bản tin rao bán của trang web nhóm đều thực kiểm tra việc lấy dữ liệu có hợp pháp hay không tại file robots.txt của trang web:

https://m.realtytrac.com/robots.txt

Ở mỗi bản tin bán nhà thành công, nhóm muốn lấy thêm một số thông tin chi tiết:

Parse HTML để lấy:

+ Taxes.

+ Price History.

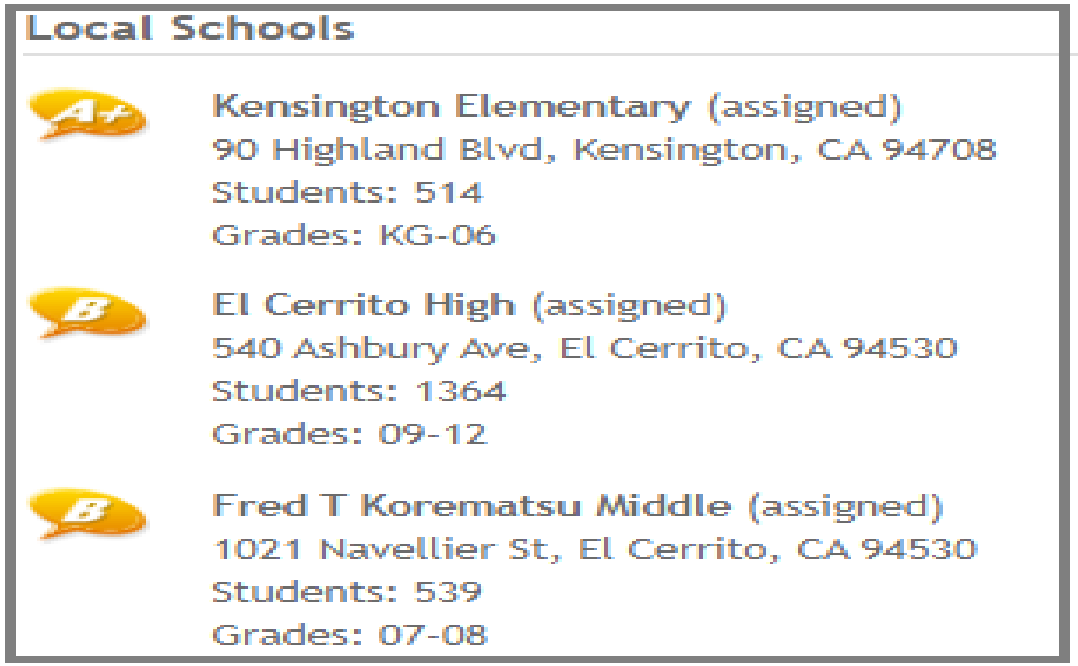
+ Number school.

Property Taxes

Land	$32,498
Improvements	$27,856
Total	$60,354
Taxes	$2,173 (3.60 %)

Price History

Sold (11/25/19)	$1,360,000
Sold (3/29/19)	$680,000



Local Schools

Kensington Elementary (assigned)
90 Highland Blvd, Kensington, CA 94708
Students: 514
Grades: KG-06

El Cerrito High (assigned)
540 Ashbury Ave, El Cerrito, CA 94530
Students: 1364
Grades: 09-12

Fred T Korematsu Middle (assigned)
1021 Navellier St, El Cerrito, CA 94530
Students: 539
Grades: 07-08

Ở mỗi lần parse tại bản tin rao bán của trang web nhóm đều thực kiểm tra việc lấy dữ liệu có hợp pháp hay không tại file robots.txt của trang web:

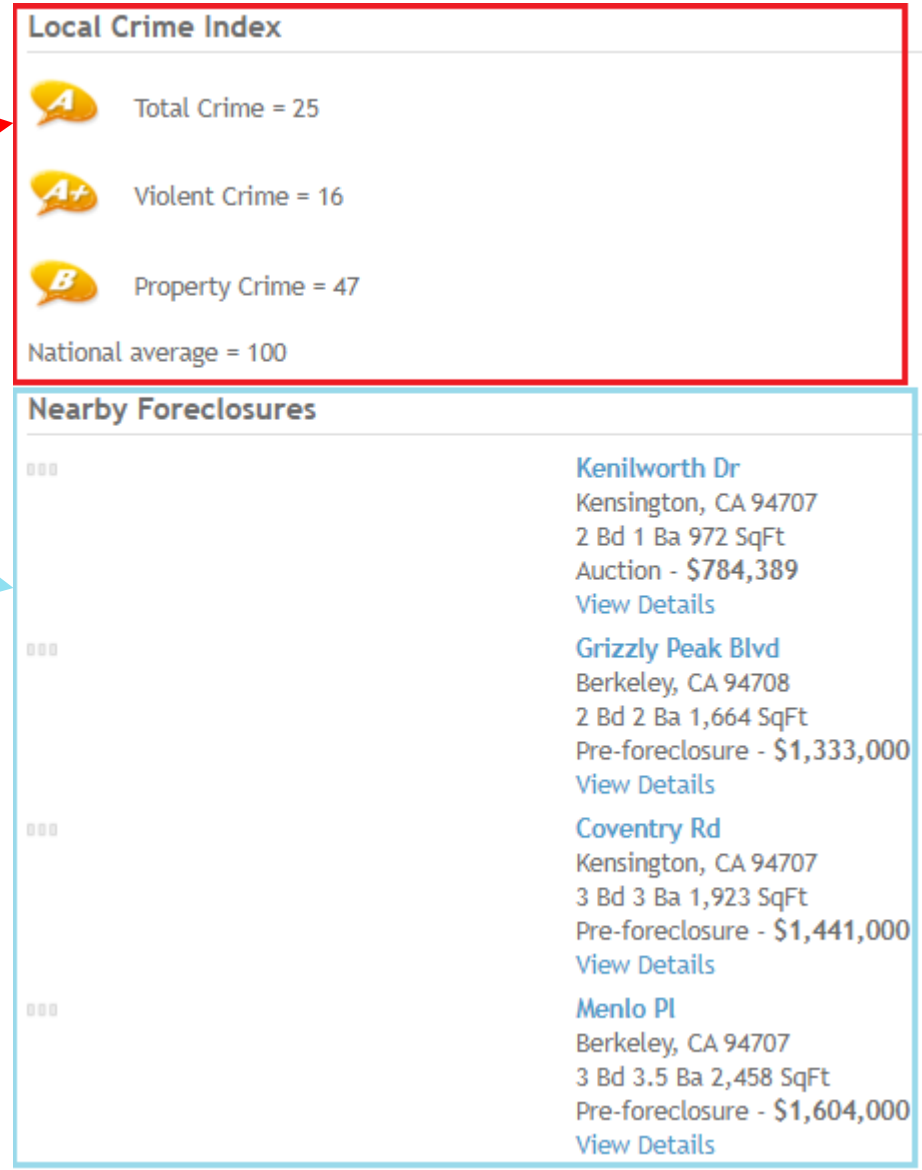
https://m.realtytrac.com/robots.txt

Ở mỗi bản tin bán nhà thành công, nhóm muốn lấy thêm một số thông tin chi tiết:

Parse HTML để lấy:

+ Info crime.

+ Number near foreclosures.



Ở mỗi lần parse tại bản tin rao bán của trang web nhóm đều thực kiểm tra việc lấy dữ liệu có hợp pháp hay không tại file robots.txt của trang web:

https://m.realtytrac.com/robots.txt

1. address_street: tên đường.
2. address_locality: tên địa phương.
3. address_region: tên vùng.
4. address_code: mã vùng.
5. date_sold: ngày bán thành công ngôi nhà.
6. mortgage: giá thuê nhà hằng tháng trước khi được bán đi.
7. info_type: loại nhà.
8. info_bedrooms: số lượng phòng ngủ.
9. info_bathrooms: số lượng phòng tắm.
10. info_size: kích thước ngôi nhà (sqft).
11.info_lot_size: kích thước lô đất (sqft).
12. info_year_build: năm ngôi nhà được xây dựng.
13. info_est_value: giá rao bán.
14. info_sold_price: giá bán thành công.
15. info_property_id: id căn nhà.
16. info_county: tên quận.
17. info_parcel_number: số bưu kiện.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9593 entries, 0 to 9592
Data columns (total 26 columns):
address_street          9593 non-null object
address_locality        9593 non-null object
address_region          9593 non-null object
address_code            9593 non-null int64
date_sold               9442 non-null object
mortgage                9481 non-null float64
info_type               9593 non-null object
info_bedrooms           7959 non-null float64
info_bathrooms          8112 non-null float64
info_size               8503 non-null float64
info_lot_size           8259 non-null float64
info_year_built         8472 non-null float64
info_est_value          822 non-null float64
info_sold_price         8659 non-null float64
info_property_id        9442 non-null float64
info_county             9593 non-null object
info_parcel_number      9593 non-null object
taxes_land              8892 non-null float64
taxes_improvements      8892 non-null float64
taxes_total             8892 non-null float64
taxes_taxes             8892 non-null object
school                  9593 non-null int64
total_crime             9593 non-null int64
violent_crime           9593 non-null int64
property_crime          9593 non-null int64
foreclosures            9593 non-null int64
dtypes: float64(12), int64(6), object(8)
memory usage: 1.9+ MB
```

Dữ liệu có 9593 dòng và 26 cột

18. axes_land: tiền đất..
19. taxes_improvements: tiền các cải tiến.
20. taxes_total: tổng của land và improvements.
21. taxes_taxes: tiền thuế từ taxes_total.
22. school: số lượng trường học gần đó.
23. total_crime: số lượng tội phạm.
24. violent_crime: số lượng tội phạm bạo lực.
25. property_crime: số lượng tội phạm về tài sản.
26. foreclosures: số lượng các căn nhà bị tịch thu gần đó.

THU THẬP DỮ LIỆU – THÔNG TIN DỮ LIỆU LẤY ĐƯỢC

Có 14 cột có giá trị thiếu là:

- date_sold thiếu 151 giá trị.
- mortgage thiếu 112 giá trị.
- info_bedrooms thiếu 1634 giá trị.
- info_bathrooms thiếu 1481 giá trị.
- info_size thiếu 1090 giá trị.
- info_lot_size thiếu 1334 giá trị.
- info_year_built thiếu 1121 giá trị.
- info_est_value thiếu 8771 giá trị.
- info_sold_price thiếu 934 giá trị.
- info_property_id thiếu 151 giá trị.
- taxes_land thiếu 701 giá trị.
- taxes_improvements thiếu 701 giá trị.
- taxes_total thiếu 701 giá trị.
- taxes_taxes thiếu 701 giá trị.

Lý do có nhiều giá trị thiếu tại các cột bên là do người muốn bán nhà lúc đăng tin không nhập đầy đủ các trường thông tin trên mà thường chỉ mô tả chi tiết tại phần description của bản tin hay vì lý do nào đó mà bản thân người bán cũng không biết rõ các thông tin trên.

Giải pháp: Sẽ bỏ đi các cột thực sự không cần thiết mà có nhiều giá trị thiếu (như info_est_value). Và phần giá trị thiếu còn lại, nhóm sẽ tiếp tục thử tìm trong phần description của mỗi bản tin rao bán từ phần dữ liệu đã lấy được ban đầu.

THU THẬP DỮ LIỆU – CÁC GIÁ TRỊ THIẾU

	address_street	address_locality	address_region	address_code	date_sold	mortgage	info_type	info_bedrooms	info_bathrooms	info_size	...	info_
0	1527 Marigold Rd	Livermore	CA	94551	11/06/2019	2089.0	Single Family Residence	2.0	2.0	866.0	...	
1	26 Highgate Rd	Kensington	CA	94707	07/10/1969	4143.0	Single Family Residence	3.0	1.5	1548.0	...	
2	28630 Barn Rock Dr	Hayward	CA	94542	11/06/2019	5360.0	Single Family Residence	4.0	2.5	2962.0	...	
3	2875 Wilson Cmn	Fremont	CA	94538	12/09/1983	3647.0	Townhouse	3.0	2.5	1748.0	...	
4	6452 Aspenwood Way	Livermore	CA	94551	11/06/2019	3820.0	Single Family Residence	4.0	2.5	1882.0	...	(
5	233 Purdue Ave	Kensington	CA	94708	11/08/2019	3906.0	Single Family Residence	2.0	1.0	1186.0	...	
6	1648 Ramblewood Way	Pleasanton	CA	94566	02/04/1965	4701.0	Single Family Residence	3.0	2.0	1675.0	...	
7	Purdue Ave	Kensington	CA	94708	11/08/2019	3906.0	MISCELLANEOUS	NaN	NaN	NaN	...	
8	1651 4th St	Livermore	CA	94550	11/06/2019	3225.0	Single Family Residence	2.0	1.0	1046.0	...	
9	190 Stanford Ave	Kensington	CA	94708	11/18/2019	5215.0	Single Family Residence	2.0	1.0	1296.0	...	

10 rows × 26 columns

THU THẬP DỮ LIỆU – 10 DÒNG ĐẦU TIÊN CỦA DỮ LIỆU ĐƯỢC THU THẬP

THANK YOU FOR WATCHING