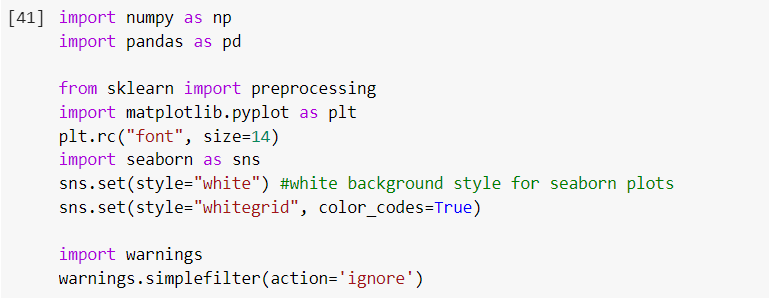
Nguyễn Vũ Quốc Đại

Mssv:17063671

Khai báo thư viện



Chuẩn bị dữ liệu, xem thông tin dữ liệu:

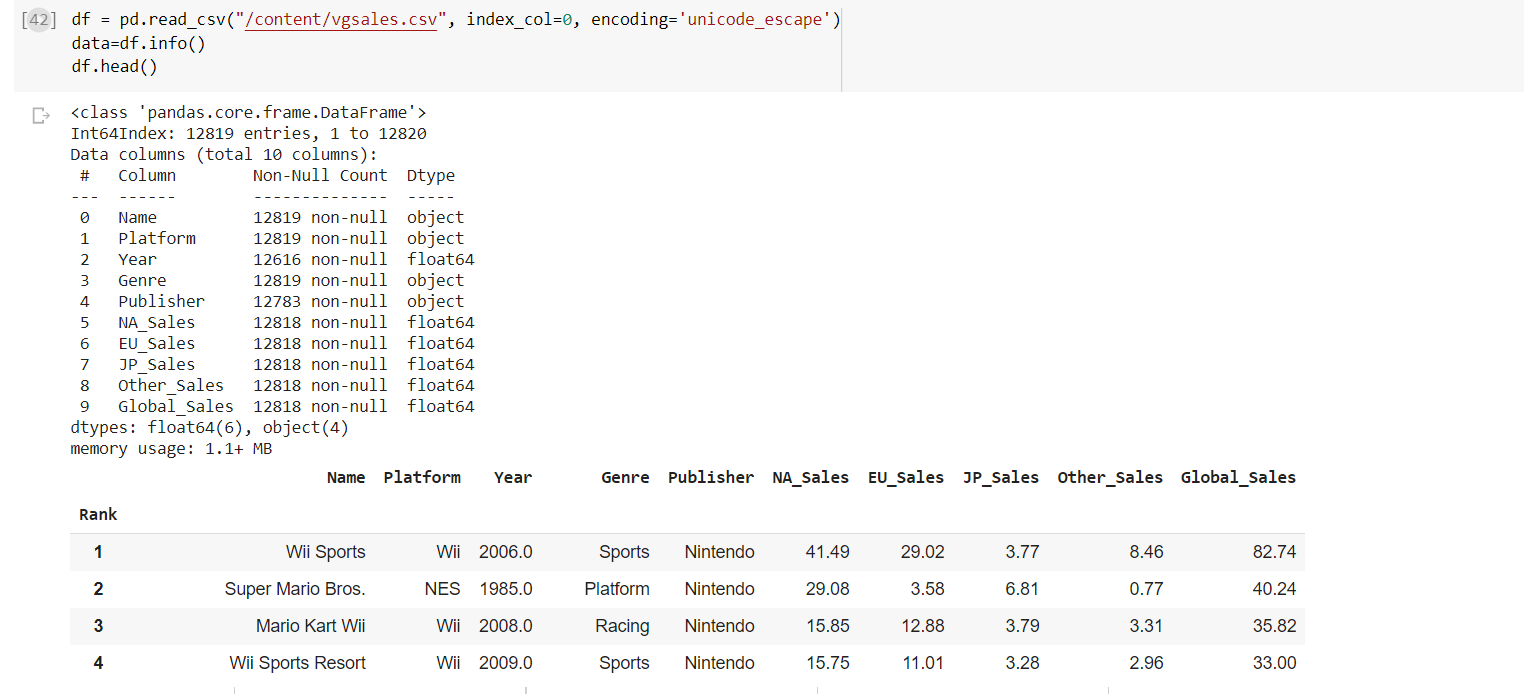
Tập dữ liệu này chứa danh sách các trò chơi điện tử có doanh số bán hàng lớn hơn 100.000 bản. Nó được tạo ra bởi một [mẩu tin lưu niệm](http://www.vgchartz.com/) của [vgchartz.com](http://www.vgchartz.com/) .

Các lĩnh vực bao gồm

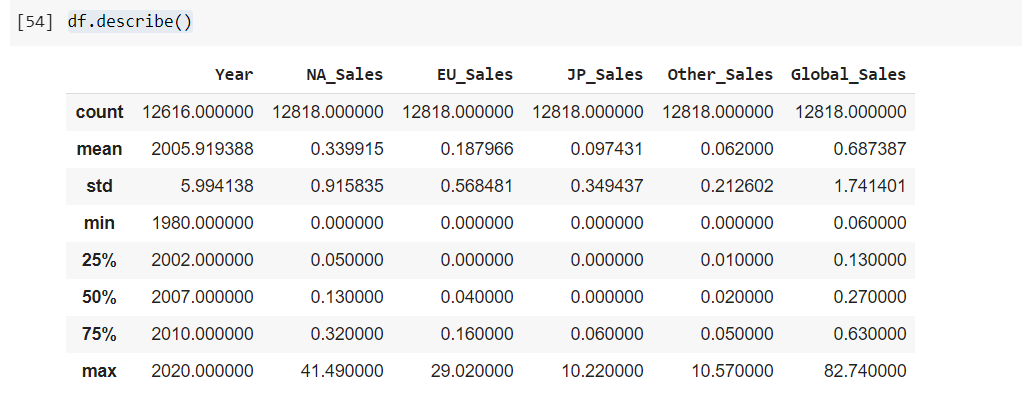
* Xếp hạng - Xếp hạng doanh số bán hàng tổng thể
* Tên - Tên trò chơi
* Nền tảng - Nền tảng phát hành trò chơi (tức là PC, PS4, v.v.)
* Năm - Năm phát hành trò chơi
* Genre - Thể loại của trò chơi
* Publisher - Nhà phát hành của trò chơi
* NA\_Sales - Bán hàng ở Bắc Mỹ (tính bằng triệu)
* EU\_Sales - Bán hàng ở Châu Âu (tính bằng triệu)
* JP\_Sales - Doanh số bán hàng tại Nhật Bản (tính bằng triệu)
* Other\_Sales - Doanh số bán hàng ở các nước còn lại trên thế giới (tính bằng triệu)
* Global\_Sales - Tổng doanh số bán hàng trên toàn thế giới.

link dữ liệu: <https://www.kaggle.com/gregorut/videogamesales?select=vgsales.csv>

Đọc file và in ra 5 dòng dữ liệu đầu tiên

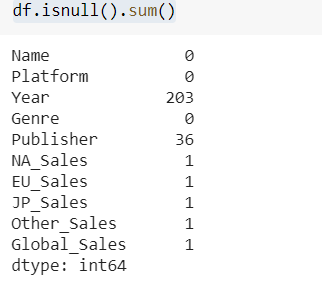


Thống kê mô tả dữ liệu



Bắt đầu xử lý dữ liệu missing

* Tìm dữ liệu bị thiếu: Có 7 cột dữ liệu bị thiếu

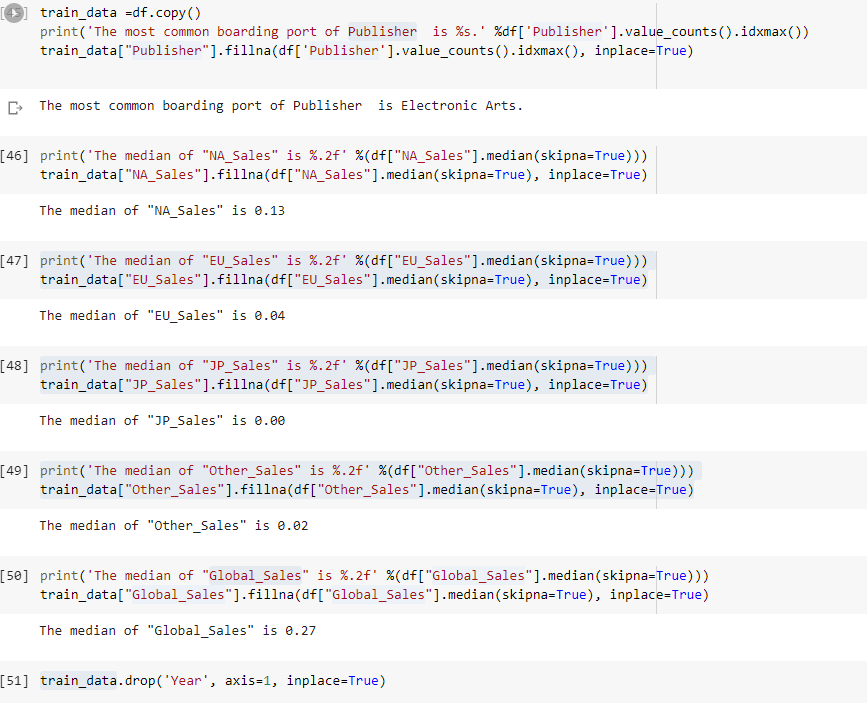


* Tính % dữ liệu bị thiếu: Ta thấy cột NA\_Sales, EU\_Sales,

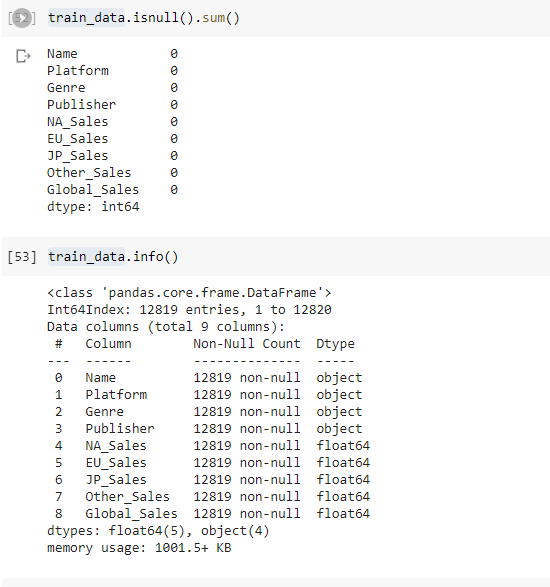
JP\_SalesOther\_Sales, Global\_Sales, dữ liệu thiếu ít nhất, cột Year dữ liệu thiếu nhiều nhất.



* Lấp đầy dữ liệu bằng giá trị được sử dụng nhiều nhất đối với cột “Publisher”. Sử dụng giá trị median để thêm vào giá trị bị thiếu, cột Year vì thiếu nhiều nên ta có thể xóa



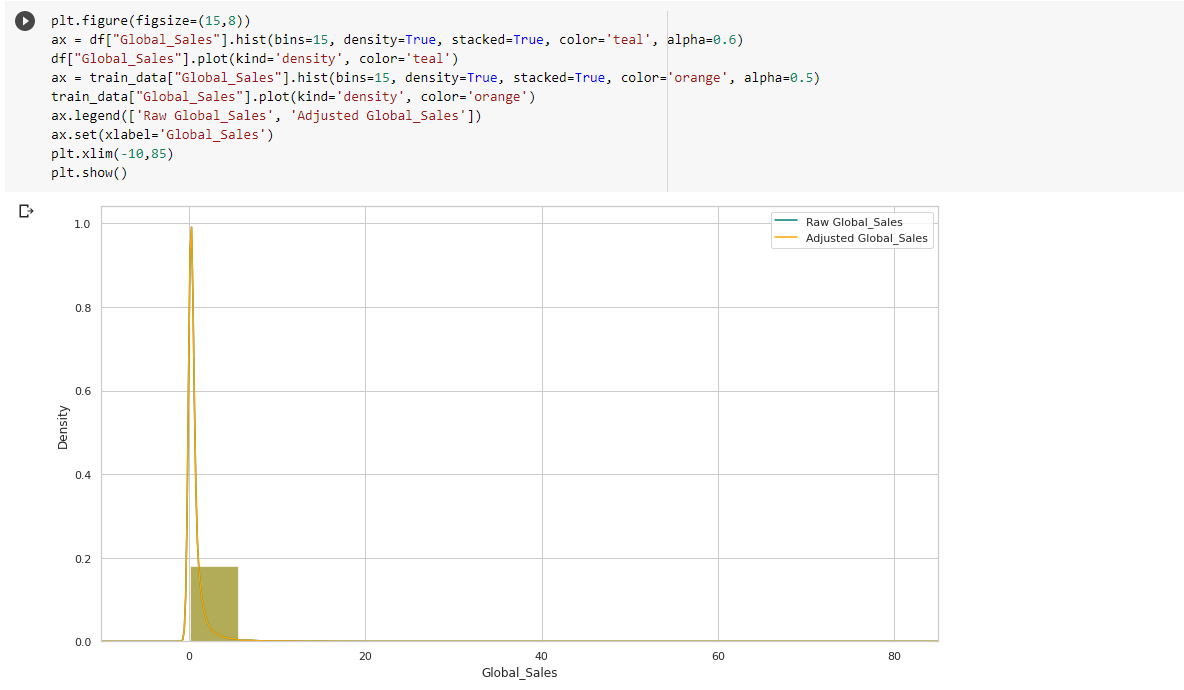
* Kiểm tra lại dữ liệu xem còn thiếu không:



* Xem lại phần đại lượng thống kê mô tả:



* Biểu diễn so sánh dữ liệu ban đầu và sau khi được xử lý: màu xanh là dữ liệu ban đầu, màu đỏ là dữ liệu sau khi thay đổi.



Biểu diễn cột “Global\_Sales” bằng biểu đồ boxplot:

