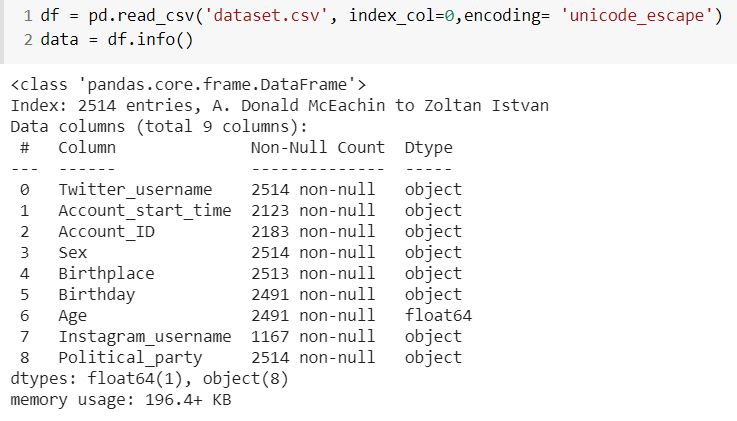
Nguyễn Vũ Quốc Đại

Mssv:17063671

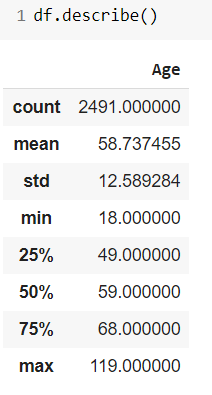
Chuẩn bị dữ liệu, xem thông tin dữ liệu:

Tập dữ liệu dựa trên tên người dùng Twitter của các chính trị gia Mỹ. Dữ liệu được trích xuất từ ​​Wikidata.

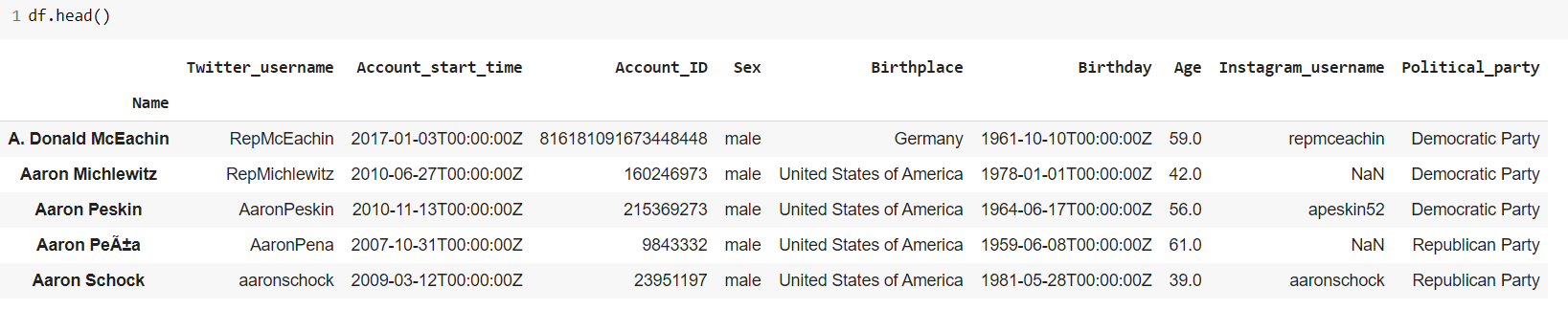
link dữ liệu <https://www.kaggle.com/mrmorj/us-politicians-twitter-dataset>



Thống kê mô tả dữ liệu

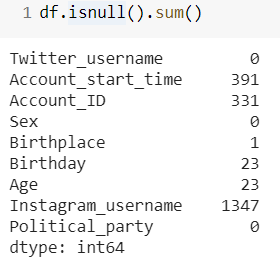


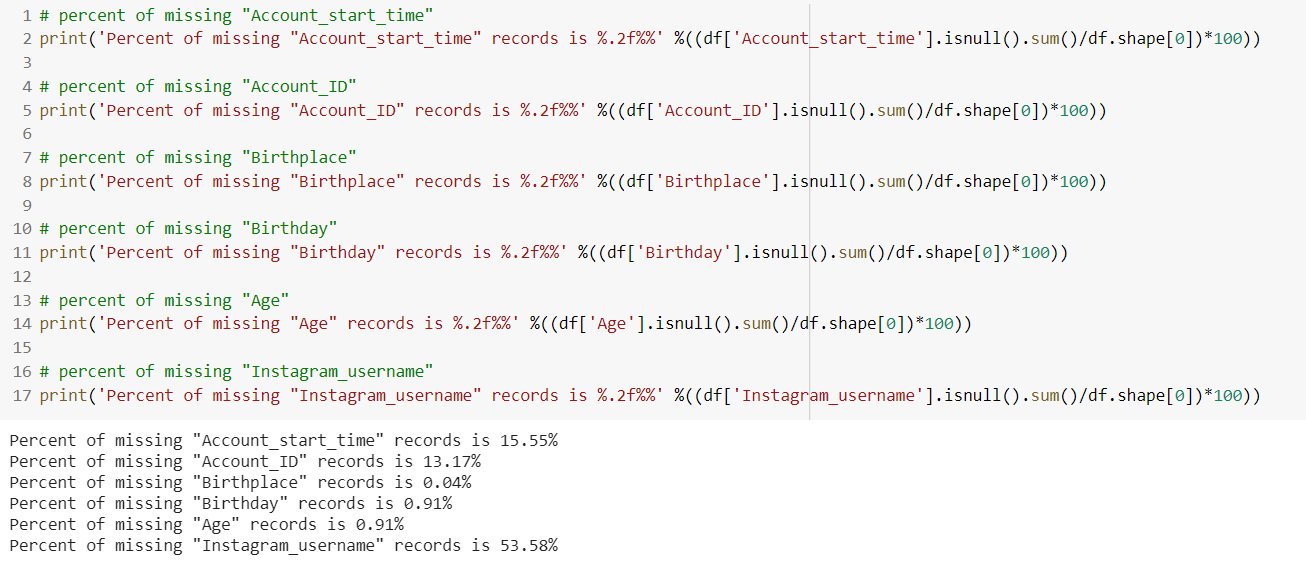
Xem 5 dữ liệu đầu tiên

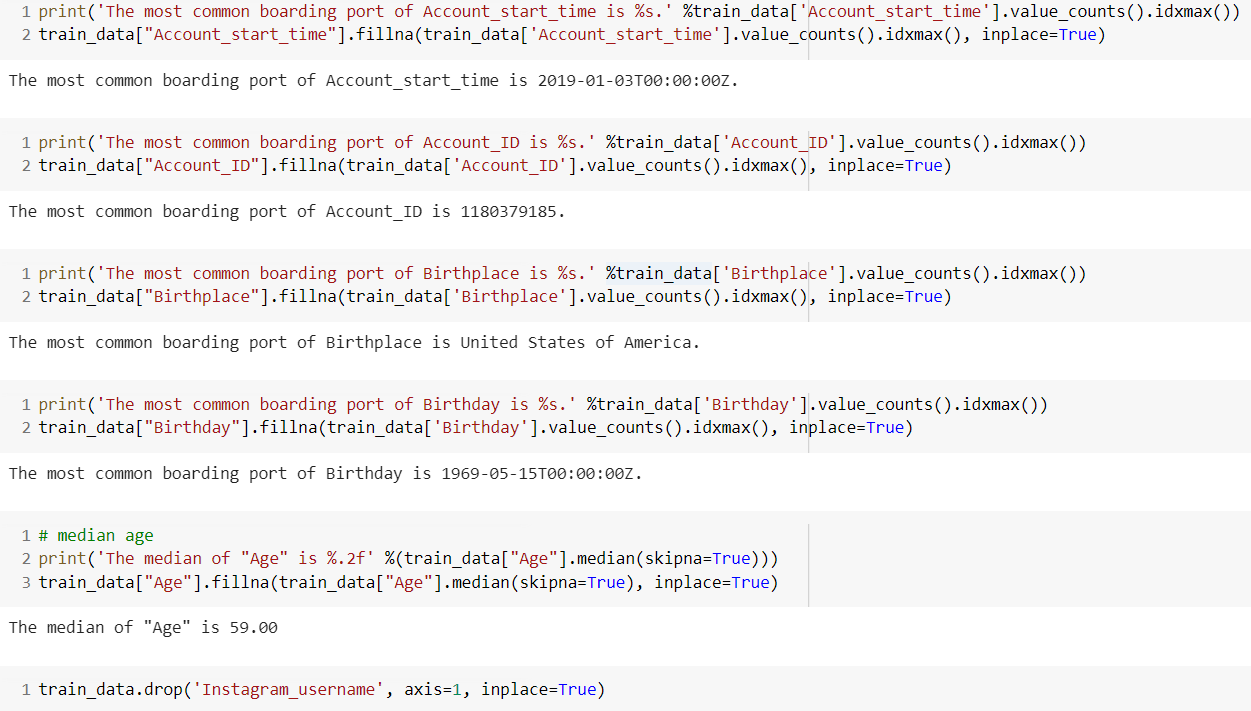


Bắt đầu xử lý dữ liệu missing

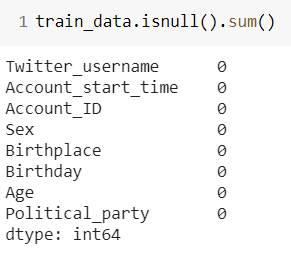
* Tìm dữ liệu bị thiếu: Có 6 cột dữ liệu bị thiếu

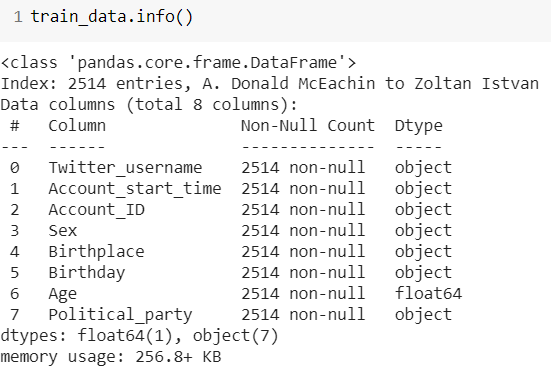


* Tính % dữ liệu bị thiếu: Ta thấy cột Birthday dữ liệu thiếu ít nhất, cột Instagram\_username dữ liệu thiếu nhiều nhất. 
* Lấp đầy dữ liệu bằng giá trị được sử dụng nhiều nhất, đối với cột “Age” thì sử dụng giá trị median để thêm vào giá trị bị thiếu, cột Instagram có thể do nhiều người không dùng Instagram nên ta có thể xóa luôn cột này.

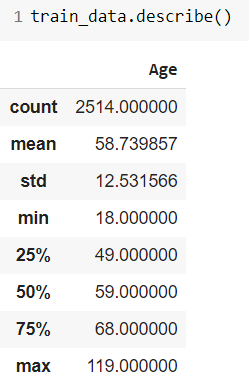


* Kiểm tra lại dữ liệu xem còn thiếu không:

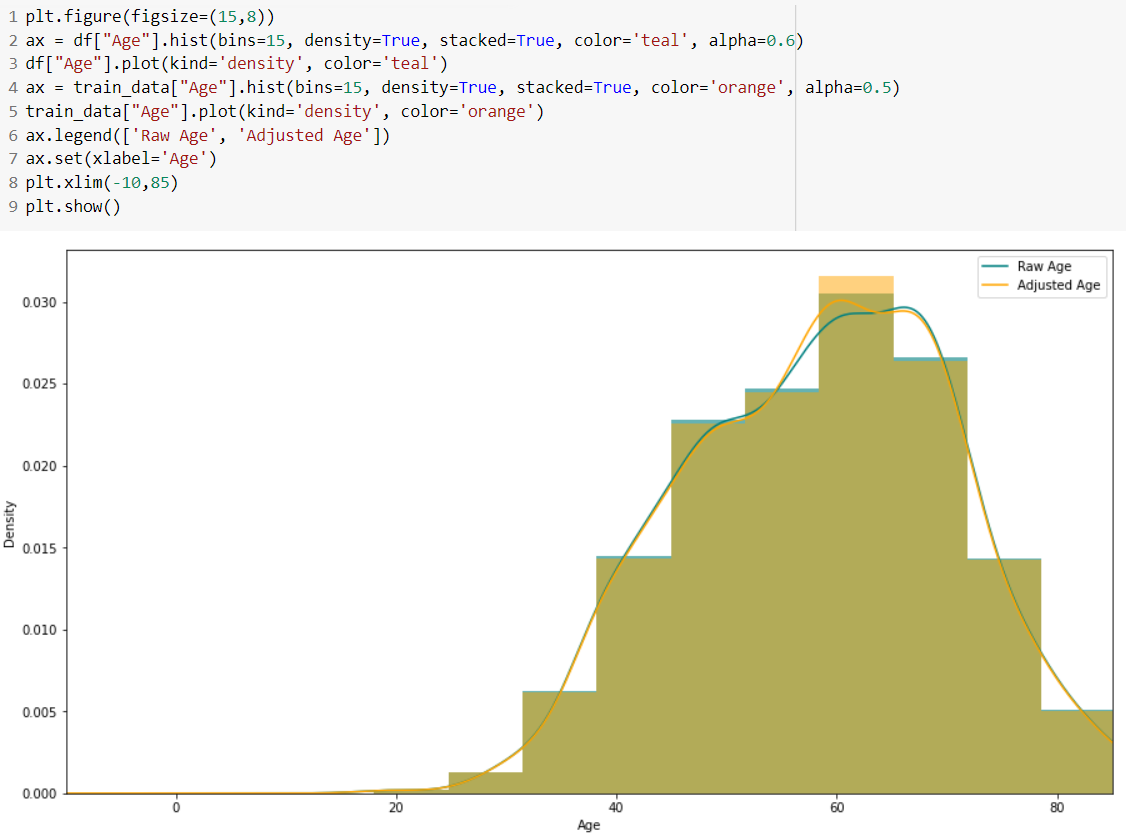




* Xem lại phần đại lượng thống kê mô tả: Tổng số lượng là 2514; giá trị trung bình là 58.739857; độ lệch chuẩn là 12.531566; giá trị nhỏ nhất là 18; giá trị lớn nhất là 119; tứ phân vị được chia ra ở các vị trí Q1=49, Q2=59, Q3=68.



* 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 “Age” bằng biểu đồ boxplot:

* Các giá trị tập trung trong khoảng từ 50 đến 70
* Giá trị lớn nhất là 119 và nhỏ nhất là 18
* Trung vị là 59

