BÁO CÁO BÀI TẬP TUẦN 6

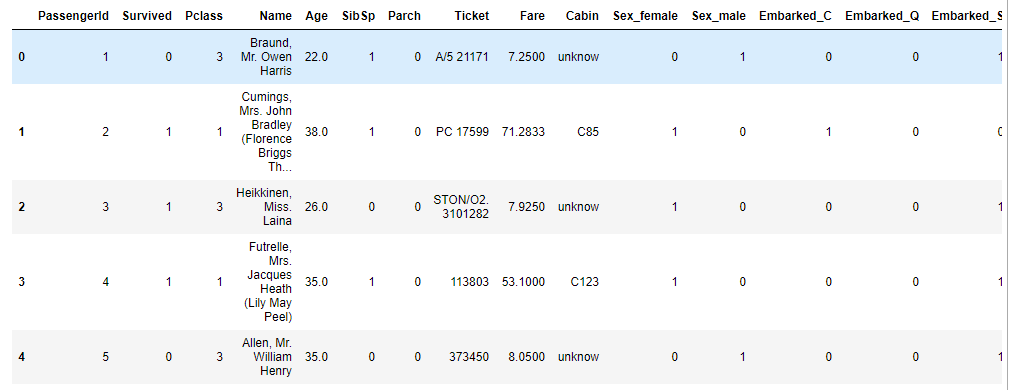
Sinh viên: Lê Kim Long

Nhóm 4

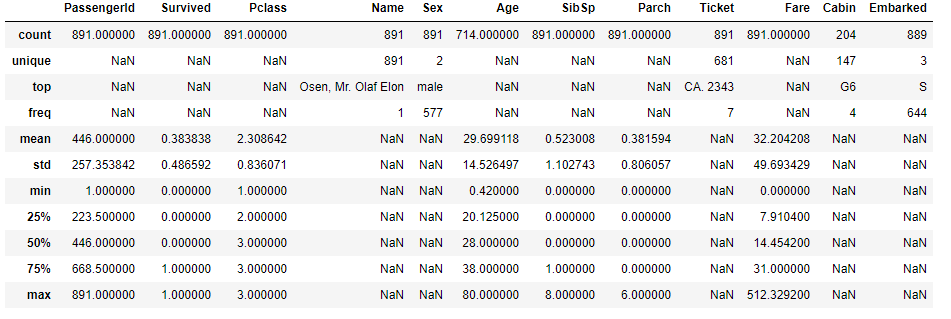
Dữ liệu: Titanic Task

1. Khảo sát và trực quan dữ liệu

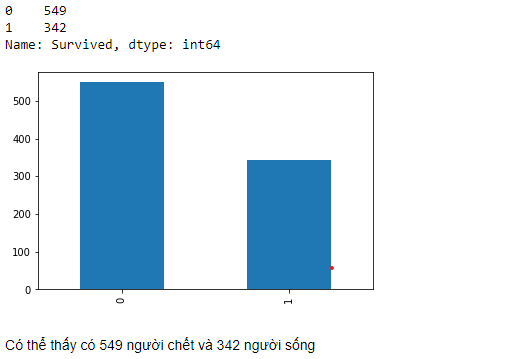
* Dữ liệu bao gồm 2 file là train.csv và test.csv. File train.csv dùng cho việc thực hiện phân lớp và sao đó test lại bằng file test.csv. File train.csv bao gồm 891 dòng và 12 cột. ( như hình)



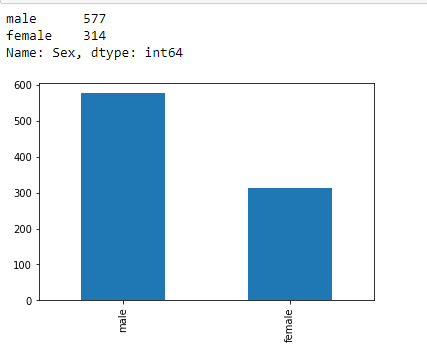
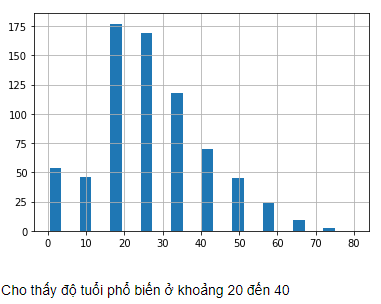
* Các mô tả và tính toán trung bình như sau

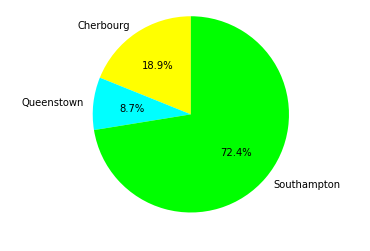


* Theo đề bảo là phân lớp sống chết và qua file train.csv chúng ta có thể tính được số người sống là số người chết

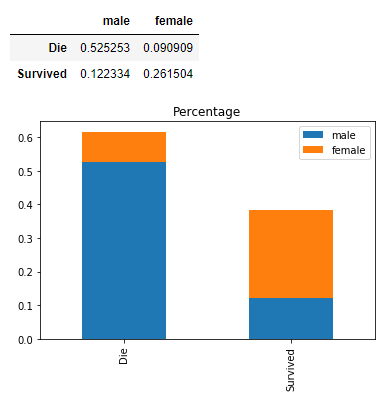


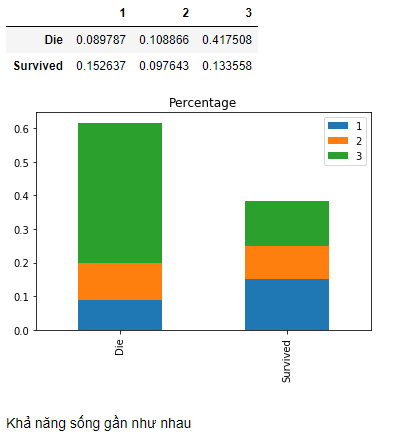
* Tiếp theo dựa vào các chỉ số có thể vẽ tiếp một số biểu đổ riêng của các cột để tiện theo rõi hơn

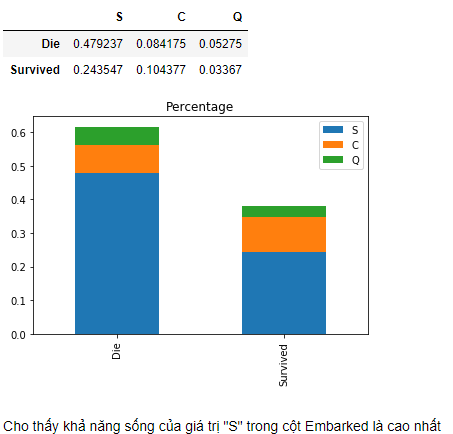




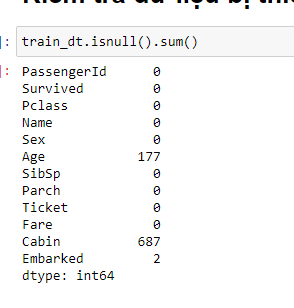
* Biểu diễn chỉ số cụ thể hơn với các giá trị của từng cột:







1. Kiểm tra vào thêm vào các phần dữ liệu bị trống mà có mục đích sử dụng đến để phân lớp.

* 
* Từ trên ta có thể thấy cột Age bị thiếu 177 giá trị cột Cabin thiếu 687 giá trị và cột Embarked là 2.
* Đối với cột Age ta có thể lấy giá trị mean của cột để thêm vào các ô chống. Cột Cabin thì ta để là unknow còn Embarked là S vì S chiếm đa số giá trị của cột.

1. Mã hoá một số cột thành số.

-Cột Sex sẽ được tách thành 2 cột là Sex\_male và Sex\_female.

Đối với cột Sex\_male những giá trị male ở Sex sẽ có giá trị là 1, female có giá trị là 0 và ngược lại ở cột Sex\_female.

* Tương tự với cột Embarked là tácht thành 3 cột Embarded\_C(Q,S)