Họ tên: Bùi Bá Minh Đăng  
Mssv: 2274802010174

**Buổi 5 Lab 4: BÀI THỰC HÀNH CHUẨN BỊ DỮ LIỆU**

PHẦN 1: DATA CLEANSING & FEATURE ENGINEERING

1. Viết hàm load\_data() để tải dữ liệu lên ứng dụng. Sau đó, hiển thị ra màn hình 10 dòng đầu tiên.

A screen shot of a computer program

AI-generated content may be incorrect.

2. Thống kê dữ liệu thiếu trên các biến số và trực quan hóa dữ liệu thiếu bằng biểu đồ (Heat map). Hãy cho nhận xét về tình trạng thiếu dữ liệu Age, Cabin và Embarked

A screenshot of a computer

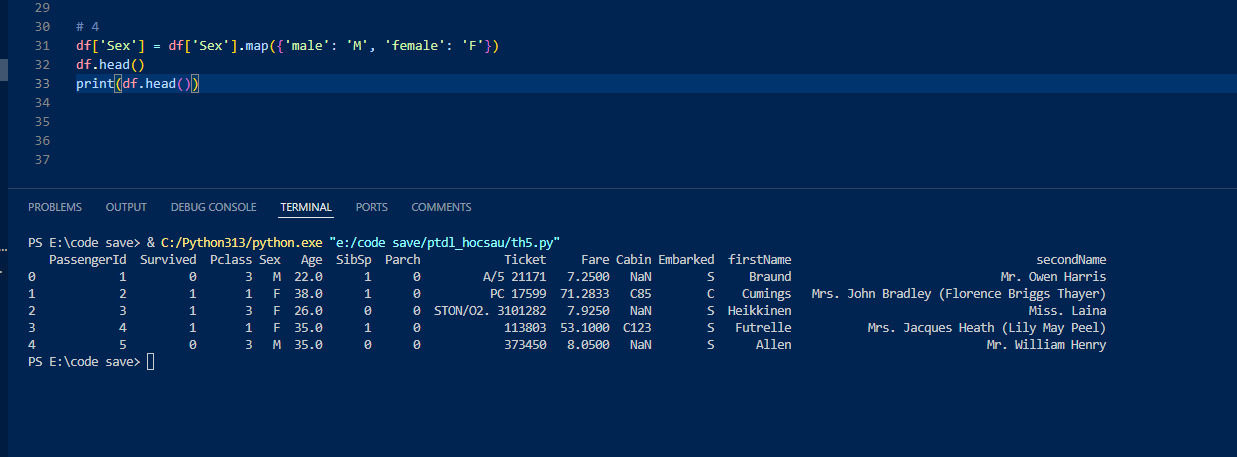
AI-generated content may be incorrect.

3. Xử lý tên cột tên Name, tách ra làm 2 cột: firstName và secondName. Lưu ý: Sau khi tách cột xong thì xóa luôn cột Name

A screen shot of a computer program

AI-generated content may be incorrect.

4. Xứ lý rút gọn kích thước dữ liệu trên cột Sex như sau: thay thế male → M và female → F



5. Xử lý dữ liệu thiếu trên biến Age bằng cách thay thế bằng giá trị trung bình tuổi: Hãy đưa ra quyết định dùng giá trị trung bình tuổi toàn bộ hành khách hay theo từng nhóm hạng vé (hạng hành khách: Pclass). Ta tiến hành làm các bước sau

A screenshot of a computer

AI-generated content may be incorrect.6. Xây dựng biến số Agegroup có thang đo thứ tự được ánh xạ theo thang đo khoảng dựa trên độ tuổi của hành khách như sau: (age =< 12] → Kid; (12, 18]: Teen, (18, 60]: Adult và ( age > 60): Older

A screenshot of a computer program

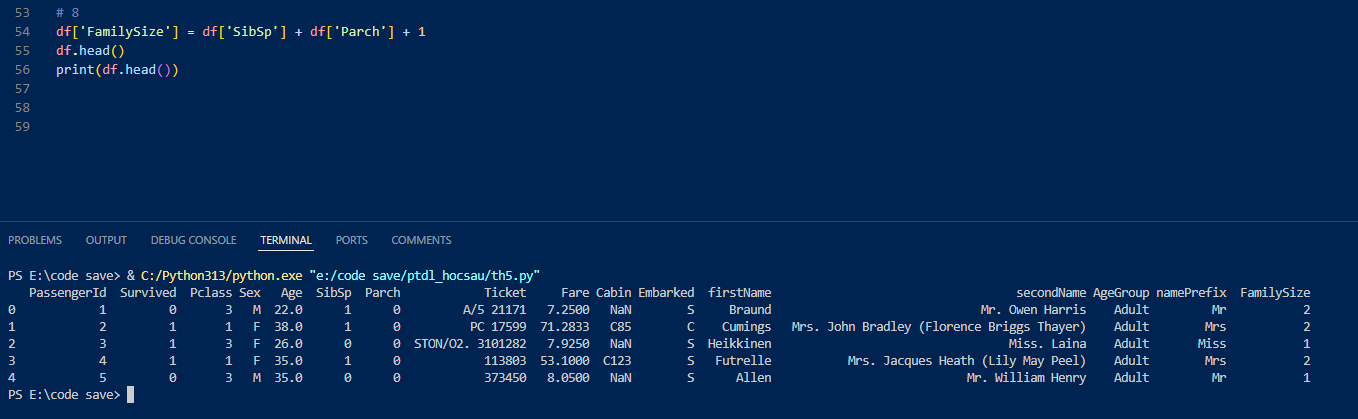
AI-generated content may be incorrect.

7. Tiến hành thêm đặc trưng về danh xưng (namePrefix) trong xã hội bằng cách tách Mr, Mrs, Miss, Master ra khỏi “secondName”

A blue screen with white text

AI-generated content may be incorrect.

8. Khai thác thêm thông tin số lượng thành viên đi theo nhóm thân quen (familySize) đối với mỗi hành khách trên chuyến hải trình; family size = 1+ SibSp + Parch



9. Tạo thêm đặc trưng ‘Alone’ để xác định hành khách đi theo nhóm hay cá nhân bằng cách dựa trên familySize như sau: Nếu familySize = 0 thì giá trị Alone = 1 và ngược lại là 0.

A blue screen with white text

AI-generated content may be incorrect.

10. Tiến hành tách loại cabin (typeCabin) mà hành khách ở để lọc và phân tích đặc tính cabin. Loại cabin được kí hiệu bởi chữ cái đầu tiên. Lưu ý: Đối với dữ liệu cabin bị thiếu thì thay thế bằng “Unknown”

A blue screen with white text

AI-generated content may be incorrect.

11. Loại bỏ dữ liệu thừa đối với các hành khách xuất hiện trong cả 2 tập dữ liệu huấn luyện (train.csv) và đánh giá (test.csv). Ưu tiên giữ lại dữ liệu trong tập huấn luyện.

A blue background with white text

AI-generated content may be incorrect.

PHẦN 2: KHAI THÁC THÔNG TIN HỮU ÍCH – EDA

12. Trực quan thông tin tương quan tỉ lệ sống sót và thiệt mạng trên từng nhóm giới tính.

A screen shot of a computer screen

AI-generated content may be incorrect.

13. Trực quan thông tin hành khách sống sót trên từng nhóm phân loại hành khách (Pclass).

A screenshot of a computer screen

AI-generated content may be incorrect.

14. Trực quan thông tin hành khách sống sót trên từng nhóm giới tính và thang đo tuổi tác

A computer screen shot of a graph

AI-generated content may be incorrect.

15. Trực quan xác suất hành khách sống sót dựa trên thông tin nhóm đi cùng

A screenshot of a computer screen

AI-generated content may be incorrect.

16. Trực quan xác suất hành khách sống sót dựa trên thông tin giá vé

A screen shot of a computer

AI-generated content may be incorrect.

17. Trực quan số lượng người thiệt mạng và sống sót theo phân lớp (Pclass) hành khách và cảng sẽ cập bến.

