**LAB THỰC HÀNH LẬP TRÌNH PHÂN TÍCH DỮ LIỆU**

**TUẦN 3: PANDAS**

**Nội dung**: Phân tích dữ liệu dulieuxettuyendaihoc.csv

***Mục tiêu***: Sinh viên đạt được kiến thức

1. Nhập liệu từ file trong pandas
2. Thao tác cơ bản với pandas
3. Thao tác gom nhóm và sử dụng các hàm tổng hợp trong pandas
4. Mô tả thống kê cho dữ liệu định tính
5. Mô tả thống kê cho dữ liệu định lượng

INSTALL PIP INSTALL PANDAS

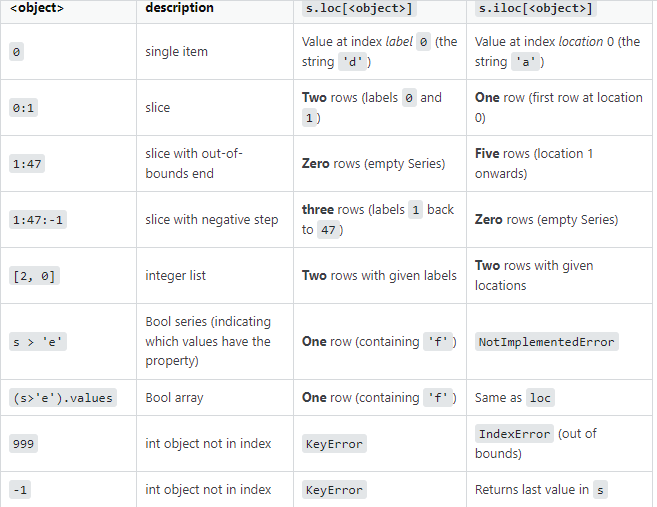
PIP INSTALL SEABORN

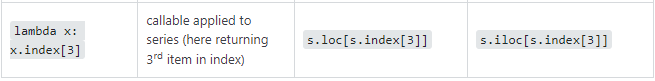
Pip install matplotlib

import pandas as pd

import matplotlib.pyplot as plt

import random





Sử dụng loc (dựa vào nhãn Lable (thứ tự từng dòng) hay Index (Tên các trường)) nhiều hơn iloc (theo dòng và cột) vì ap dung cho ca hai

Lable: đại lượng đặc trưng cho tên của hàng đó

**BÀI THỰC HÀNH TUẦN 3 - PANDAS**

1. Đọc File với các file trong thư mục dữ liệu. Ứng với mỗi file sinh viên thực hiện các câu sau. Hiển thị toàn bộ dữ liệu của file dữ liệu đã đọc
2. Đọc file với chỉ định không có header, dòng header của chúng ta đã biến thành 1 bản ghi dữ liệu. Hiển thị toàn bộ dữ liệu của file dữ liệu đã đọc
3. Đọc File với các tuỳ chọn mặc định. Hiển thị 5 dòng dữ liệu đầu tiên
4. Đọc File với các tuỳ chọn mặc định. Hiển thị 5 dòng dữ liệu cuối cùng
5. Chuyển kiểu dữ liệu cho 1 cột nào đó
6. Xem chiều dài của df, tương đương shape[0]

print('Len:', len(df))

1. Xem thông tin dataframe vừa đọc được

df.info()

1. Xem kích thước của dataframe

print('Shape:', df.shape)

1. Hiển thị dữ liệu của cột thứ 3

df[‘tên cột’]

1. Hiển thị dữ liệu của cột 1,2,3,5,6

df[['Tên cột 1', 'Tên cột 2', 'Tên cột 3']]

1. Hiển thị 5 dòng dữ liệu đầu tiên gồm các cột 1,2,3,5,6

df[['Tên cột 1', 'Tên cột 2', 'Tên cột 3']].head[5]

1. Hiển thị 5 dòng dữ liệu đầu tiên theo chỉ số

df[0:5]

1. Hiển thị 5 dòng dữ liệu đầu tiên theo chỉ số gồm các cột 1,2,3,5,6

df[['Tên cột 1', 'Tên cột 2', 'Tên cột 3']][:5]

1. Loại bỏ các dòng trùng nhau

df.drop\_duplicates(inplace=True)

1. Loại bỏ các dòng trống có dữ liệu của 1 cột là trống

df['ADDRESSLINE2'].isna().sum()

1. Loại bỏ các dòng của 1 cột có giá trị là không biết (“unknown”)

df["ADDRESSLINE2"].fillna('Unknown',inplace=True)

df['ADDRESSLINE2'].isna().sum()

1. Lấy dữ liệu của 1 cột theo dạng chuỗi

df["QUANTITYORDERED"]

1. Lấy dữ liệu của 1 cột về một mảng

df["QUANTITYORDERED"].values

1. Lấy dữ liệu của 1 cột về một mảng

df["QUANTITYORDERED"].values

1. Lấy dữ liệu từ dòng số 4 đến dòng 9

df[4:10]

1. Đọc dữ liệu từ dòng 4 đến dòng 9

df.loc[4:10]

df.iloc[4:10]

1. Lấy thông tin tại dòng có chỉ số là 2

df.loc[2]

1. Lấy thông tin từ dòng 4 đến dòng 10 của một số cột

df.loc[4:10,['QUANTITYORDERED','SALES']]

1. Lấy thông tin dòng 2 đến dòng 9, từ cột 4 đến cột 7

df.iloc[2:9,4:7]

1. Lấy dữ liệu tại chỉ số (index) là 2

df.iloc[2]

1. Lấy dữ liệu từ dòng đầu tiên đến dòng 9 dùng iloc

df.iloc[:10]

1. Lấy dữ liệu từ dòng đầu tiên đến dòng 9 gồm các cột 4 đến cột 7 dùng iloc

df.iloc[2:9,4:7]

1. Sắp xếp dữ liệu theo 1 cột tăng dần

df.sort\_values(by='SALES',ascending=True)

1. Sắp xếp dữ liệu theo nhiều tiêu chí

df.sort\_values(by=['QUANTITYORDERED','PRICEEACH'],ascending=[True,False])

1. Lọc dữ liệu theo 1 điều kiện

df[df[‘SALES']>5000]

df.loc[df['SALES']>5000]

1. Lọc dữ liệu theo nhiều điều kiện

df[(df['SALES']>5000) & (df['QUANTITYORDERED']>40)]

1. Lọc giá trị và gán điều kiện dùng loc

df.loc[df['PRICEEACH']>=65,'FLAG']='EXPENSIVE'

df.loc[df['PRICEEACH']<65,'FLAG']='CHEAP'

#df['FLAG']

df[['PRICEEACH','FLAG']]#[:50]

1. Viết hàm trả về giá trị có nhiều điều kiện và áp dụng hàm gán trị trả về cho 1 cột

def foo(x):

    if x<10:

        return 'BAD'

    elif x>=10 and x<50:

        return 'GOOD'

    else:

        return 'EXCELLENT'

df['WORTH']=df[['QUANTITYORDERED']].applymap(foo)

df[['QUANTITYORDERED','WORTH']]

1. Ánh xạ giá trị tới 1 cột

dict\_map ={1:'Qui\_1',2:'Qui\_2',3:'Qui\_3',4:'Qui\_4'}

df['QTR\_ID']=df['QTR\_ID'].map(dict\_map)

df['QTR\_ID']

1. Lấy những dòng dữ liệu bằng 1 điều kiện nào đó

df[['QUANTITYORDERED','PRICEEACH']].loc[df['YEAR\_ID']==2003]

1. Hiển thị các bản ghi có cột kiểu số hơn 25

df[df[‘age’]<25]

Tuoitre =df[df[‘age’]<25]

Tuoitre[:5]

1. Sinh viên tự nghĩ ra các câu hỏi đọc dữ liệu theo các điều kiện và thực hiện lại các câu hòi đó. Yêu cầu sinh viên ghi lại và lệnh thực hiện các câu hòi đó.
2. Thực hiện 1 ví dụ để lấy giá trị của một cột trả về dưới dạng numpy array trong thư viện pandas python.
3. Sử dụng thư viện random để sinh ngẫu nhiên một list năm sinh và thêm vào dataframe.
4. Thêm 1 cột vào file dữ liệu
5. Thêm 1 cột vào dữ liệu theo tiêu chí nếu điều kiện thoả thì giá trị mặc định là True, ngược lại là False.
6. Tạo 1 cột mới có giá trị rỗng
7. Thêm 1 bản ghi trong dataframe
8. Sửa giá trị của cột
9. Xóa cột trong dataframe
10. Xóa bản ghi theo chỉ số
11. Sử dụng hàm describe() để thống kê dữ liệu
12. Xem thống kê trên từng cột
13. Vẽ đồ thị xem phân bổ giá trị của 1 trường trong dataframe
14. **Tạo mới dataframe từ các python list**
15. Sắp xếp dataframe
16. Nối 2 dataframe (Lỗi)
17. Xáo trộn các bản ghi trong dataframe
18. Lưu dataframe về file csv
19. Tối ưu bộ nhớ khi dùng pandas
20. Tạo 1 file chương trình hiện các menu gồm các mục trên và mục cuối là thoát. Người dùng chọn chức năng nào trong menu thì chương trình sẽ thực hiện chức năng tương ứng.