

**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT HƯNG YÊN**  
**KHOA CÔNG NGHỆ THÔNG TIN**

----- oOo -----



# **BÀI TẬP THỰC HÀNH**

# **LẬP TRÌNH PYTHON NÂNG CAO**

**TRÌNH ĐỘ: ĐẠI HỌC CHÍNH QUY**

**NGÀNH ĐÀO TẠO: KHOA HỌC MÁY TÍNH**

**Hưng Yên – Tháng 11 năm 2021**

## BÀI THỰC HÀNH SỐ 2: THAO TÁC VỚI DỮ LIỆU SỬ DỤNG PANDAS

### A. MỤC TIÊU BÀI THỰC HÀNH

Sau bài thực hành này sinh viên có thể:

- Hiểu được cách tạo một Pandas Series, một Pandas DataFrame trong Python
- Biết cách truy xuất các phần tử trong một Pandas Series
- Biết cách truy xuất dữ liệu theo hàng, cột của Pandas DataFrame
- Sử dụng thành thạo các hàm trong Pandas Series và Pandas DataFrame
- Vận dụng các kỹ thuật xử lý dữ liệu với Pandas để phân tích dữ liệu trong các bài toán thực tế

### B. ĐIỀU KIỆN THỰC HÀNH

Với đặc thù của môn Lập trình Python nâng cao, mục này sẽ liệt kê một số công cụ sử dụng để làm bài thực hành. Trong bài thực hành này, sinh viên cần kiểm tra và chắc chắn các phần mềm sau trên máy tính còn hoạt động tốt:

1. Anacoda phiên bản 3.0 trở lên
2. Jupyter Notebook

### C. TÀI NGUYÊN THAM CHIẾU

Để hoàn thành tốt bài thực hành này sinh viên nên tham khảo các tài nguyên sau:

STT	Tên tài nguyên	Mô tả tài nguyên
1	Practice 02.pdf	Tài liệu hướng dẫn thực hành bài số 1
2	Lesson 03 – Working with Pandas.pdf	Slide bài giảng về thao tác dữ liệu với Pandas

### D. YÊU CẦU BÀI THỰC HÀNH

**Bài 1.** Thực hiện các thao tác cơ bản với Pandas Series trong Python

- a) Tạo hai Pandas Series **a** và **b** tương ứng với 2 mảng dữ liệu sau:  
[1, 2, 3, 4, 5] và [6, 7, 8, 9, 10]
- b) Viết chương trình tính tổng, hiệu, tích, thương của **a** và **b**

c) Viết chương trình so sánh **a** và **b**

**Bài 2.** Cho dữ liệu điểm học phần của một lớp chứa trong một Pandas Series, **s**, như sau: [8.5, 9.0, 7.0, 3.0, 'CĐG', 6.0, 7.5, 'CĐG', 8.0]

- a) Chuyển kiểu dữ liệu trong **s** thành dạng số chứa trong mảng **a**
- b) Thay thế các giá trị không phải số trong **a** bằng giá trị **0**
- c) Tìm giá trị lớn nhất, nhỏ nhất của **a**
- d) Tính tổng, giá trị trung bình (mean), giá trị trung vị (median) của **a**

**Bài 3:** Sắp xếp và trích chọn dữ liệu trong Pandas Series

- a) Tạo một Pandas Series **d** gồm 20 giá trị random từ 0 đến 1
- b) Sắp xếp **d** theo giá trị tăng dần
- c) Sắp xếp **d** theo giá trị giảm dần
- d) Trích chọn từ **d** các phần tử lớn hơn hoặc bằng 0.8
- e) Trích chọn từ **d** các phần tử nằm trong đoạn [0.5,0.79]

**Bài 4:** Thực hiện các thao tác cơ bản với Pandas DataFrame trong Python

- a) Tạo một Pandas DataFrame **d** để lưu trữ dữ liệu thông tin chi tiết của các hóa đơn như sau:

OrderID	ProductID	ProductName	Quantity	Price	Discount
101	LAC	Acer Laptop	1	18900000	0.5
101	MLT	Logitech Mouse	2	300000	0.3
102	LAC	Acer Laptop	3	18900000	0.5
102	HPA	Panasonic Headphone	1	1000000	0.2
103	MMS	Microsoft Mouse	1	250000	0.6

- b) Hiển thị dữ liệu toàn bộ dữ liệu trong DataFrame ra màn hình
- c) Hiển thị dữ liệu các trường “OrderID, ProductID, ProductName” trong DataFrame ra màn hình
- d) Truy xuất lấy về dữ liệu của cột ProductID
- e) Truy xuất lấy về dữ liệu của một mã hóa đơn cụ thể
- f) Truy xuất lấy về các mặt hàng đã bán với chiết khấu  $\geq 0.5$  (50%)
- g) Tính tổng doanh thu trên toàn bộ sản phẩm bán được
- h) Tính tổng doanh thu trên mỗi sản phẩm, hiển thị ra màn hình với các trường như sau:  
ProductID, ProductName, Revenue (Doanh thu)

**Bài 5:** Sử dụng Pandas DataFrame trong Python để phân tích dữ liệu của một Shop thời trang với 44,446 sản phẩm được lưu trữ trong tệp **styles.csv**

a) Đọc toàn bộ dữ liệu trong tệp **styles.csv** ra DataFrame và hiển thị ra màn hình

	id	gender	masterCategory	subCategory	articleType	baseColour	season	year	usage	productDisplayName
0	15970	Men	Apparel	Topwear	Shirts	Navy Blue	Fall	2011.0	Casual	Turtle Check Men Navy Blue Shirt
1	39386	Men	Apparel	Bottomwear	Jeans	Blue	Summer	2012.0	Casual	Peter England Men Party Blue Jeans
2	59263	Women	Accessories	Watches	Watches	Silver	Winter	2016.0	Casual	Titan Women Silver Watch
3	21379	Men	Apparel	Bottomwear	Track Pants	Black	Fall	2011.0	Casual	Manchester United Men Solid Black Track Pants
4	53759	Men	Apparel	Topwear	Tshirts	Grey	Summer	2012.0	Casual	Puma Men Grey T-shirt
...	...	...	...	...	...	...	...	...	...	...

b) Hiển thị ra các sản phẩm của năm 2019

c) Hiển thị ra các sản phẩm mùa “Winter” năm 2019

d) Hiển thị ra các sản phẩm áo “Tshirts” của “Men” năm 2019

e) Thống kê số lượng sản phẩm:

(1) Theo giới tính

(2) Theo năm

(3) Theo loại sản phẩm (articleType)

f) Thống kê số lượng các sản phẩm mỗi mùa của năm 2019 với 2 cột:

Season, Quantity

## E. HƯỚNG DẪN THỰC HIỆN

Sinh viên tạo tệp **Practice02\_HoVaTen.ipynb** trên Jupyter Notebook và thực hiện viết mã lệnh để giải quyết các bài tập thực hành.

### Bài 1:

- Sinh viên sử dụng kiến thức về Pandas Series để làm bài tập

### Bài 2:

- Sinh viên sử dụng kiến thức về Pandas Series để làm bài tập

### Bài 3:

- Sinh viên sử dụng kiến thức về Pandas Series để làm bài tập
  - a) Sử dụng phương thức rand của numpy để tạo dữ liệu ngẫu nhiên

### Bài 4:

- Sinh viên sử dụng kiến thức về Pandas DataFrame để làm bài tập

h) Sử dụng hàm **sum**

g) Sử dụng hàm **groupby** và hàm **sum**

**Bài 5:**

a) Sử dụng phương thức `read_csv` để đọc dữ liệu

f) Sử dụng hàm **groupby** và hàm **sum**