SỞ GIÁO DỤC VÀ ĐÀO TẠO
THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG THPT CHUYÊN TRẦN ĐẠI NGHĨA

ĐỀ CHÍNH THỨC

(Đề gồm 08 trang)

ĐỀ KIỂM TRA GIỮA – HKI
NĂM HỌC 2024 - 2025
MÔN: TIN HỌC (CHUYÊN) – LỚP 12
Ngày kiểm tra: 25/10/2024
Thời gian làm bài thực hành: 45 phút (*Không tính thời gian phát đề*)

HƯỚNG DẪN LÀM BÀI VÀ NỘP BÀI

- Bài làm thực hiện bằng ngôn ngữ Python và các thư viện liên quan.
- Môi trường làm bài có thể là Anaconda, Google colab hoặc Visual Studio Code
- Bài làm thực hiện theo nhóm, lưu tên bài là: <Tên nhóm>_KTGHKI.ipynb
- Cuối giờ nộp vào thư mục drive do thầy add nhóm trưởng vào.
- Dataset sử dụng trong đề thi được gửi kèm theo đề.

Phần Thực hành:

Bài 1: Dùng Pandas và Matplot để tạo danh sách thành viên của nhóm theo dạng sau, phải fix đúng định dạng theo dạng bảng: (*1điểm*)

	SO TT TRONG LOP	NAME
1	18	Nguyễn Thị Phương Thảo
2	19	Phạm Văn Minh Phúc
3	20	Nguyễn Minh Hoàng
4	21	Phạm Thời Ngô Huy

Bài 2: Sử dụng Padas và Numpy (*3 điểm*)

2.1. Tạo Dataframe theo mẫu say và tính trung bình lương theo nghề nghiệp, kết quả hiện thị như mẫu: (0.6)

```
                    Name Occupation  Salary
0       Nguyễn Minh Hoàng   engineer   60000
3  Nguyễn Thị Phương Thảo     doctor   60000
4   Nguyễn Trọng Thái Sơn     doctor   65000
1        Phạm Thời Ngô Huy     doctor   70000
2      Phạm Văn Minh Phúc   engineer   50000
Occupation
doctor      65000.0
engineer    55000.0
Name: Salary, dtype: float64
```

2.2. Thêm hai dòng dữ liệu vào Dataframe trên nhưng điền cho lương là NaN, sau đó thay giá trị ‘0’ vào các vị trí NaN đó. (0.6)

Ví dụ:

Dữ liệu trước và sau như sau:

Trước:	**Sau:**

	name	occ	salary
0	Vinay	engineer	NaN
1	Kushal	doctor	70000.0
2	Aman	engineer	NaN
3	Rahul	doctor	60000.0
4	Ramesh	doctor	65000.0

	name	occ	salary
0	Vinay	engineer	0.0
1	Kushal	doctor	70000.0
2	Aman	engineer	0.0
3	Rahul	doctor	60000.0
4	Ramesh	doctor	65000.0

2.3. Apply function (0.6)

Hãy tạo ra một Dataframe có hai cột company name và profit như sau:

	cname	profit
0	Shyam & Co.	-10000
1	Ramlal & Bros.	10000
2	Sharma Enterprises	-5000
3	Verma Furnitures	15000
4	Rahul Stores	20000

Sau đó áp dụng function trên cột profit để được kết quả như sau:

	cname	profit
0	Shyam & Co.	False
1	Ramlal & Bros.	True
2	Sharma Enterprises	False
3	Verma Furnitures	True
4	Rahul Stores	True

2.4. Ghép Dataframe dựa vào cột chung (0.6)

Hãy tạo ra hai Dataframe như sau:

Dataframe 1:

	eid	ename	stipend
0	1	Sid	10000
1	2	Ramesh	10000
2	3	Ron	5000
3	4	Harry	15000

Dataframe 2:

	eid	position
0	1	employee
1	2	employee
2	3	intern
3	4	senior_employee

Sau đó ghép lại dựa vào cột eid để được kết quả sau:

	eid	ename	stipend	position
0	1	Sid	10000	employee
1	2	Ramesh	10000	employee
2	3	Ron	5000	intern
3	4	Harry	15000	senior employee

2.5. Thống kê các thông số của Dataframe (0.6)

Hãy tạo một Dataframe như sau:

	eid	ename	stipend	position
0	1	Sid	10000	employee
1	2	Ramesh	10000	employee
2	3	Ron	5000	intern
3	4	Harry	15000	senior_employee

Sau đó hãy viết lệnh để có được kết quả thống kê sau của các trường số:

	eid	stipend
count	4.000000	4.000000
mean	2.500000	10000.000000
std	1.290994	4082.482905
min	1.000000	5000.000000
25%	1.750000	8750.000000
50%	2.500000	10000.000000
75%	3.250000	11250.000000
max	4.000000	15000.000000

Bài 3: Đọc và xử lý dữ liệu từ file CSV, TSV (*3 điểm*)

3.1. Đọc dữ liệu từ file IMDB -Movie - Data .csv sau đó thực hiện: (0.375)

- Hiển thị 5 dòng đầu tiên

- Hiển thị các thông tin cơ bản của dữ liệu.

- Đưa ra nhận một số nhận xét về dữ liệu.

3.2. Hãy tách cột Genr thành series (0.375)

Kết quả như sau:

```
0       Action,Adventure,Sci-Fi
1      Adventure,Mystery,Sci-Fi
2               Horror,Thriller
3       Animation,Comedy,Family
4      Action,Adventure,Fantasy
                ...
995         Crime,Drama,Mystery
996                      Horror
997         Drama,Music,Romance
998            Adventure,Comedy
999       Comedy,Family,Fantasy
Name: Genre, Length: 1000, dtype: object
```

3.3. Tạo một DataFrame mới gồm các trường Title, Rating, Revenue(Miilions) và lấy dữ liệu từ dòng 10 tới 14. Kết quả có dạng như sau: (0.375)

	Title	Rating	Revenue (Millions)
10	Fantastic Beasts and Where to Find Them	7.5	234.02
11	Hidden Figures	7.8	169.27
12	Rogue One	7.9	532.17
13	Moana	7.7	248.75
14	Colossal	6.4	2.87

3.4. các bộ phim từ 2010 tới 2015, với rating nhỏ hơn 6.0 nhưng lại có doanh thu thuộc top 5% trên toàn bộ dataset, kết quả có dạng như sau: **(0.375)**

	Rank	Title	Genre	Description	Director	Actors	Year	Runtime (Minutes)	Rating	Votes
941	942	The Twilight Saga: Eclipse	Adventure,Drama,Fantasy	As a string of mysterious killings grips Seatt...	David Slade	Kristen Stewart, Robert Pattinson, Taylor Laut...	2010	124	4.9	192740

3.5. Tính số rating trung b.nh mà các đạo diễn đạt được, kết quả có dạng như sau: (0.375)

Director	Rating
Aamir Khan	8.5
Abdellatif Kechiche	7.8
Adam Leon	6.5
Adam McKay	7.0
Adam Shankman	6.3

3.6. Dựa trên kết quả của câu 3.5, em hãy đưa ra top 5 đạo diễn có rating trung bình cao nhất. Kết quả có dạng như sau: (0.375)

Director	Rating
Nitesh Tiwari	8.80
Christopher Nolan	8.68
Olivier Nakache	8.60
Makoto Shinkai	8.60
Aamir Khan	8.50

3.7. Thống kê số dòng mất dữ liệu theo từng cột, kết quả thống kê có dạng sau: (0.375)

```
Rank                    0
Title                   0
Genre                   0
Description             0
Director                0
Actors                  0
Year                    0
Runtime (Minutes)       0
Rating                  0
Votes                   0
Revenue (Millions)    128
Metascore              64
dtype: int64
```

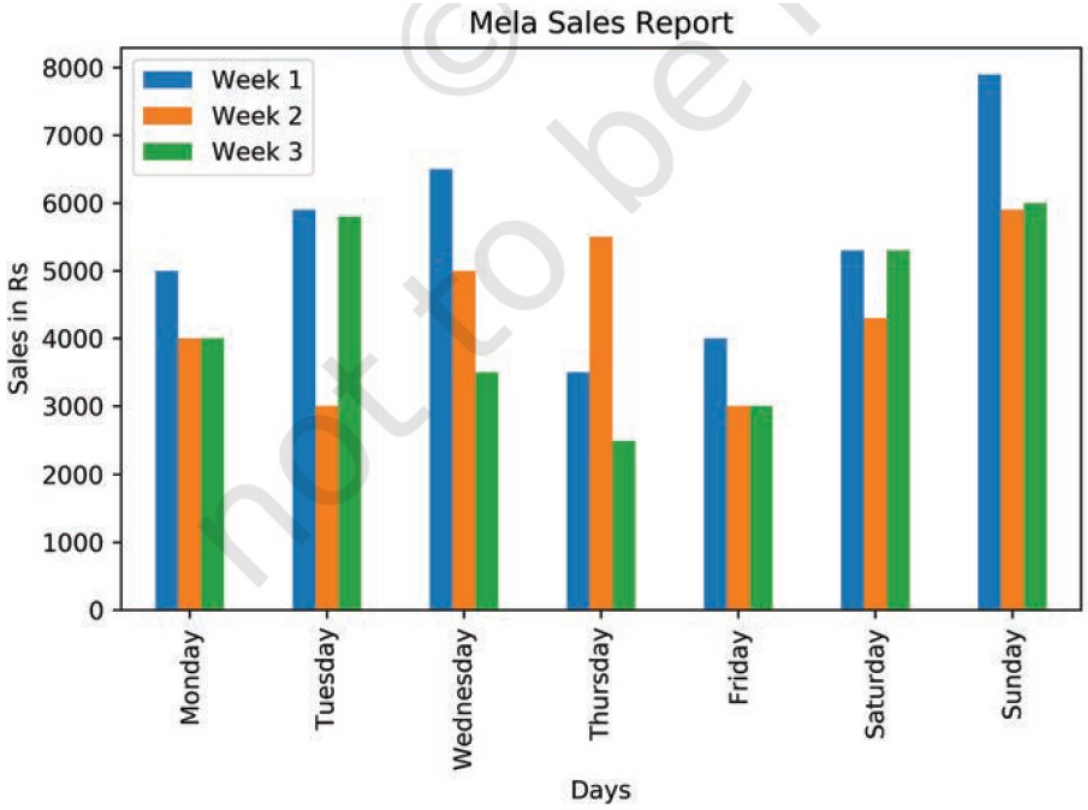
Sau đó, hãy thay hết các vị trí mất dữ liệu của cột Revenue (Millions) bằng giá trị trung bình của cột đó, sau đó xuất lại bảng thống kê trên một lần nữa.

3.8. Hãy thêm một cột mới với tên "Rating_category" để xếp hạng các phim dựa trên giá trị rating theo ba mức ['Good', 'Average', 'Bad']. Kết quả có dạng như sau: (0.375)

	Title	Director	Rating	Rating_category
0	Guardians of the Galaxy	James Gunn	8.1	Good
1	Prometheus	Ridley Scott	7.0	Average
2	Split	M. Night Shyamalan	7.3	Average
3	Sing	Christophe Lourdelet	7.2	Average
4	Suicide Squad	David Ayer	6.2	Average

Bài 4. Visualize data với Matplot (3đ)

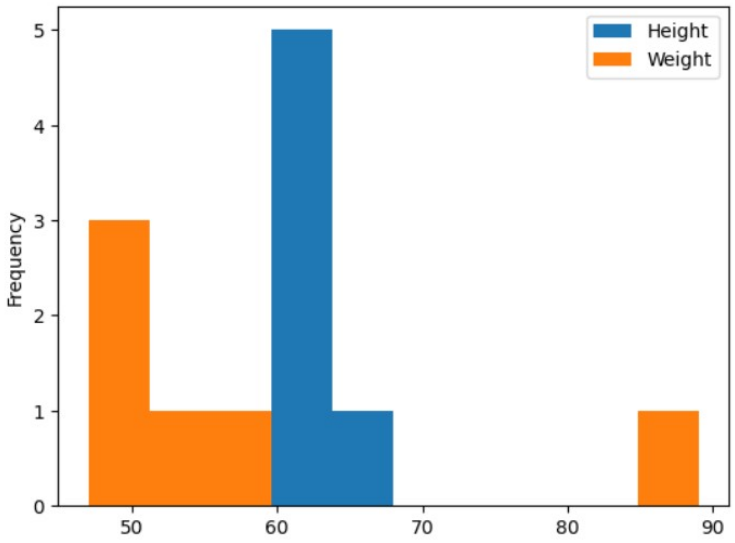
4.1. Sử dụng Dataset MelaSale.csv để vẽ biểu đồ thể hiện doanh thu theo ngày của từng tuần. Biểu đồ có dạng sau: (0.6)



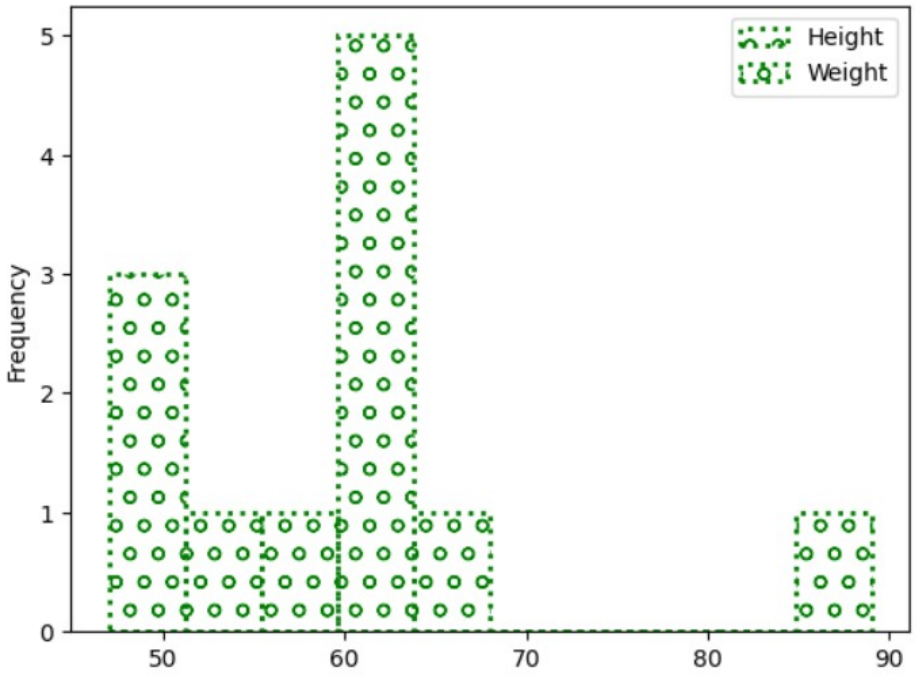
4.2. Tạo ra một DataFrame có dạng như sau: (0.6)

	Name	Height	Weight
0	Arnav	60	47
1	Sheela	61	89
2	Azhar	63	52
3	Bincy	65	58
4	Yash	61	50
5	Nazar	60	47

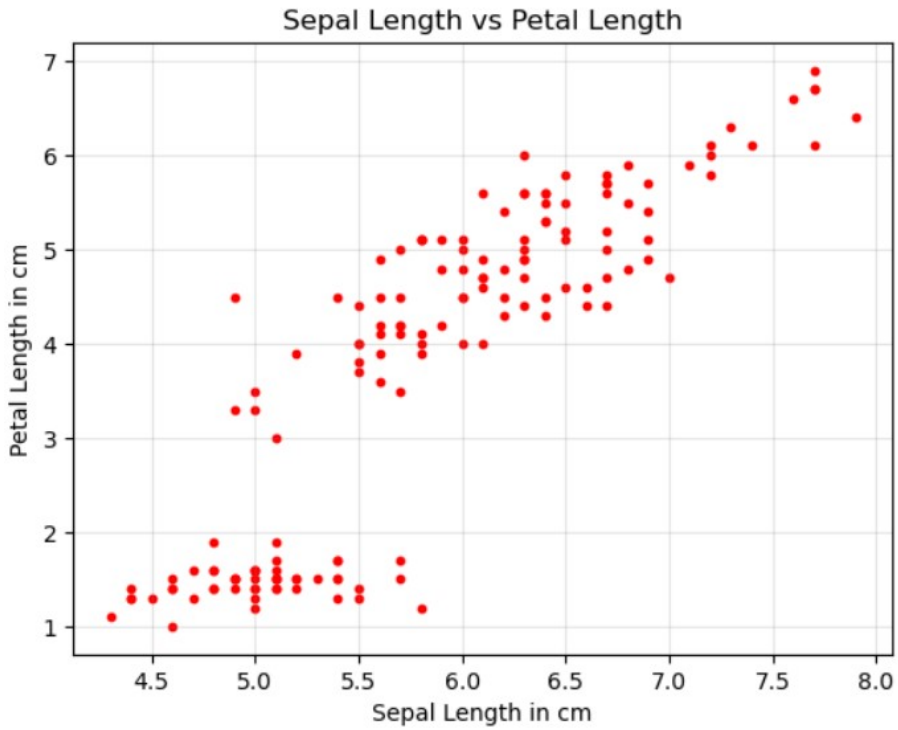
Sau đó vẽ biểu đồ theo hình dạng sau:



4.3. Hãy Custom lại biểu đồ ở câu 4.2 thành hình dạng sau: (0.6)



4.4. Sử dụng DataSet Iris.csv để vẽ biểu đồ tương quan của đài hoa và cánh hoa, biểu đồ có dạng như sau: (0.6)



Hãy nhận xét sơ lượt về mối tương quan này (nhận xét bằng cách print("…."), nhận xét của em nằm trong dấu "…")

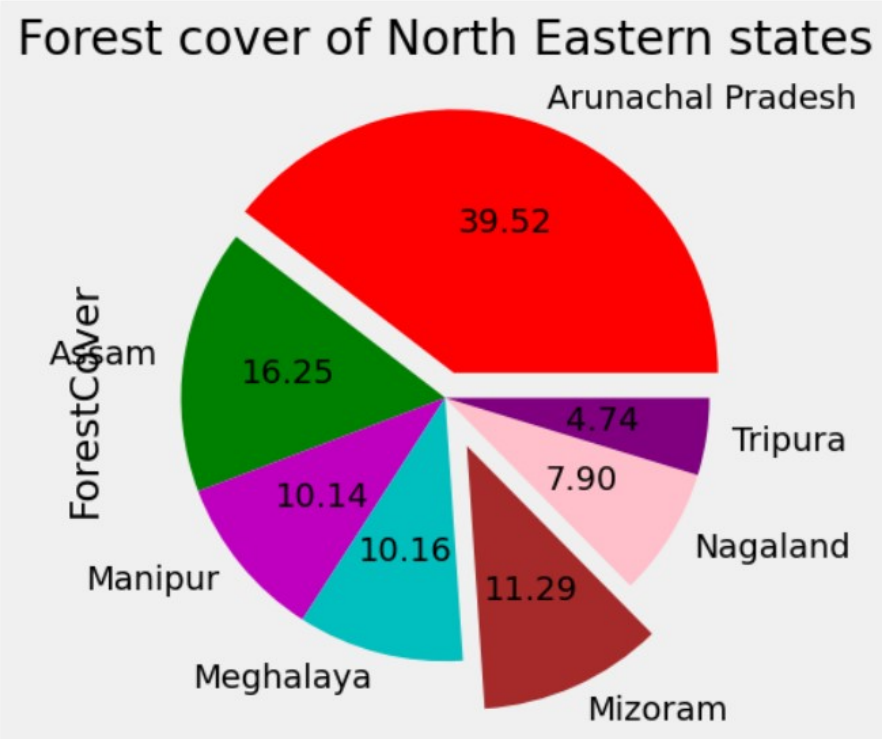
4.5. Sau đây là bảng thống kê diện tích và diện tích rừng bao phủ của một số bang tại một quốc gia: (0.6)

State	GeoArea	FrestCover
Arunachal Pradesh	83743	67353
Assam	78438	27692
Manipur	22327	17280
Meghalaya	22429	17321
Mizoram	21081	19240
Nagaland	16579	13464
Tripura	10486	8073

Hãy tạo ra DataFrame như sau:

	GeoArea	ForestCover
Arunachal Pradesh	83743	67353
Assam	78438	27692
Manipur	22327	17280
Meghalaya	22429	17321
Mizoram	21081	19240
Nagaland	16579	13464
Tripura	10486	8073

Sau đó vẽ biểu đồ thể hiện phần trăm diện tích rừng của từng bang so với tổng diện tích rừng của bảy bang. Biểu đồ có hình dạng như sau:



--- HẾT ---

PHẦN THI VẤN ĐÁP

Phần này sẽ được hỏi trực tiếp trong lúc làm bài hoặc sau khi làm xong, mỗi học sinh được hỏi một hoặc một vài câu liên quan trực tiếp đến các nhiệm vụ thực hành ở trên.