Đề thi:

DATA MANIPULATION AND VISUALIZATION WITH PYTHON

Thời gian: 120 phút

**** Học viên tạo 1 thư mục là **MD21_K288_HoVaTen**, lưu tất cả bài làm vào để nộp chấm điểm ****

**** Học viên được sử dụng tài liệu ****

Chú ý, với mỗi câu:

- Học viên cần kiểm tra xem dữ liệu có bị thiếu (NaN, null, hoặc để trống) hay không, nếu có thì cần chuẩn hóa trước khi làm bài.
- Cần hiển thị thông tin chung của dữ liệu bằng cách dùng shape, head(), tail(), info()… để có cái nhìn ban đầu về dữ liệu.
- Lần lượt thực hiện các bước làm bài như đã được hướng dẫn làm bài tập trong lớp.
- Mỗi câu là 1 file viết trên Jupyter Notebook, các yêu cầu nhận xét kết quả trong từng câu được viết trong cell dưới định dạng Markdown.

1. Numpy Array (1.5 điểm)

- Cho dữ liệu data với các giá trị như sau :
  data = [30, 22, 47, 20, 46, 47, 37, 24, 40, 12, 43, 45, 44, 48, 24, 30, 15, 28, 21, 29, 17, 24, 41, 33, 23]
- **Yêu cầu**: sử dụng thư viện Numpy thực hiện các yêu cầu sau :
  - Tạo mảng arr từ data và in ra kết quả như sau (0.25 điểm):

```
Mảng arr: [30 22 47 20 46 47 37 24 40 12 43 45 44 48 24 30 15 28 21 29 17 24 41 33 23]

5 phần tử đầu tiên của mảng: [30 22 47 20 46]
5 phần tử cuối cùng của mảng: [17 24 41 33 23]
```

  - Từ mảng 1 chiều arr, chuyển đổi thành mảng 2 chiều arr_2d có kích thước 5x5 và in mảng arr_2d ra màn hình như sau (0.25 điểm).

```
Mảng 2 chiều:
[[30 22 47 20 46]
 [47 37 24 40 12]
 [43 45 44 48 24]
 [30 15 28 21 29]
 [17 24 41 33 23]]
```

  - Lọc và in ra các phần tử có giá trị là số chẵn và số lẻ trong arr_2d (0.25 điểm)

```
Các phần tử có giá trị là số chẵn:

array([30, 22, 20, 46, 24, 40, 12, 44, 48, 24, 30, 28, 24])

Các phần tử có giá trị là số lẻ:

array([47, 47, 37, 43, 45, 15, 21, 29, 17, 41, 33, 23])
```

  - Đếm số phần tử chẵn và số phần tử lẻ có trong mảng arr_2d (0.25 điểm)

```
Số lượng phần tử có giá trị chẵn: 13
Số lượng phần tử có giá trị lẻ: 12
```

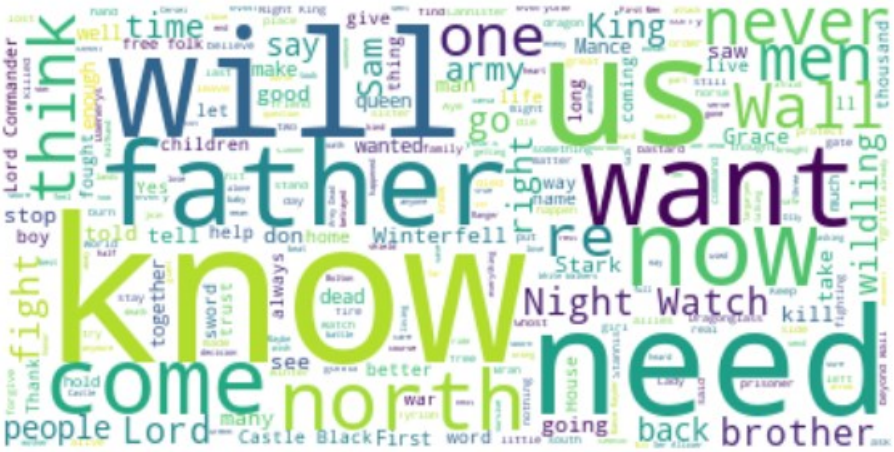
  - Thay thế các giá trị chẵn trong mảng arr_2d bằng 0 và thay thế các giá trị lẻ trong mảng arr_2d bằng 1 và in ra kết quả (0.5 điểm)

```
Mảng 2 chiều sau khi thay thế:

array([[0, 0, 1, 0, 0],
       [1, 1, 0, 0, 0],
       [1, 1, 0, 0, 0],
       [0, 1, 0, 1, 1],
       [1, 0, 1, 1, 1]])
```

2. Game of Thones dialogue (1.5 điểm)

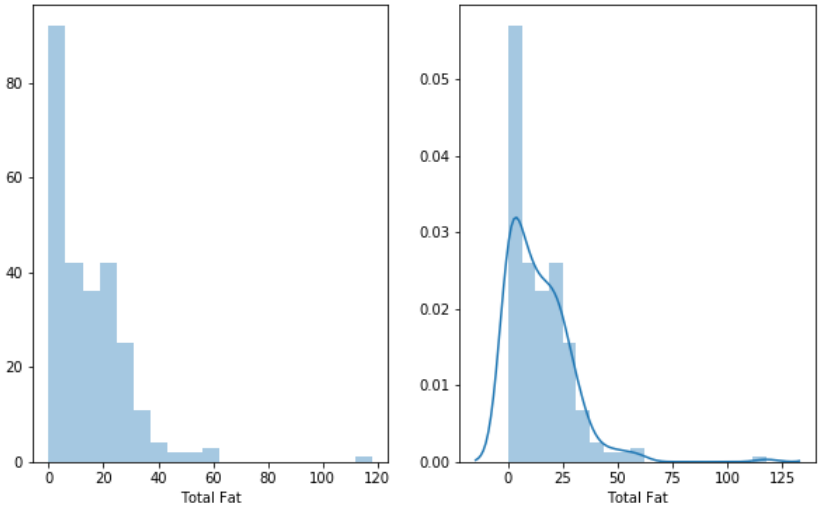
- Cho dữ liệu **Game_of_Thrones_jon_snow_data.csv** thực hiện các yêu cầu sau :
  1. Đọc dữ liệu và thực hiện chuẩn hóa (loại bỏ phần văn bản subject trước dấu : và các ký tự đặc biệt như \r, \n) (0.5 điểm)
  2. Tạo biểu đồ Wordcloud có kết quả gợi ý như sau : (0.5 điểm)



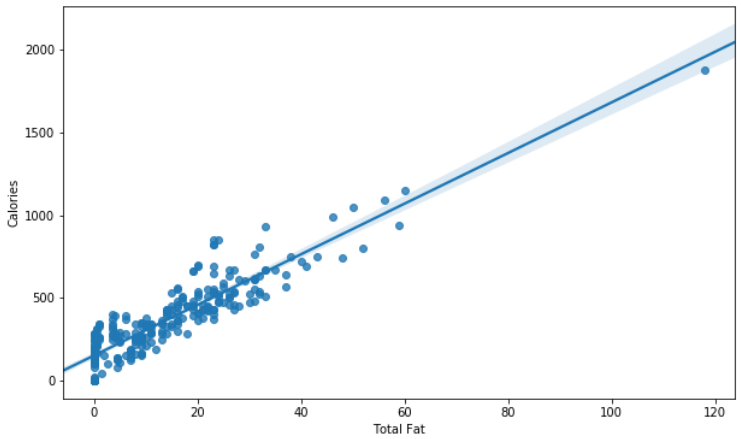
  3. Cho tập tin hình ảnh **jon-snow.jpg**, hãy tạo biểu đồ có kết quả gợi ý như hình sau : (0.5 điểm)

3. McDonald's menu: (4 điểm)

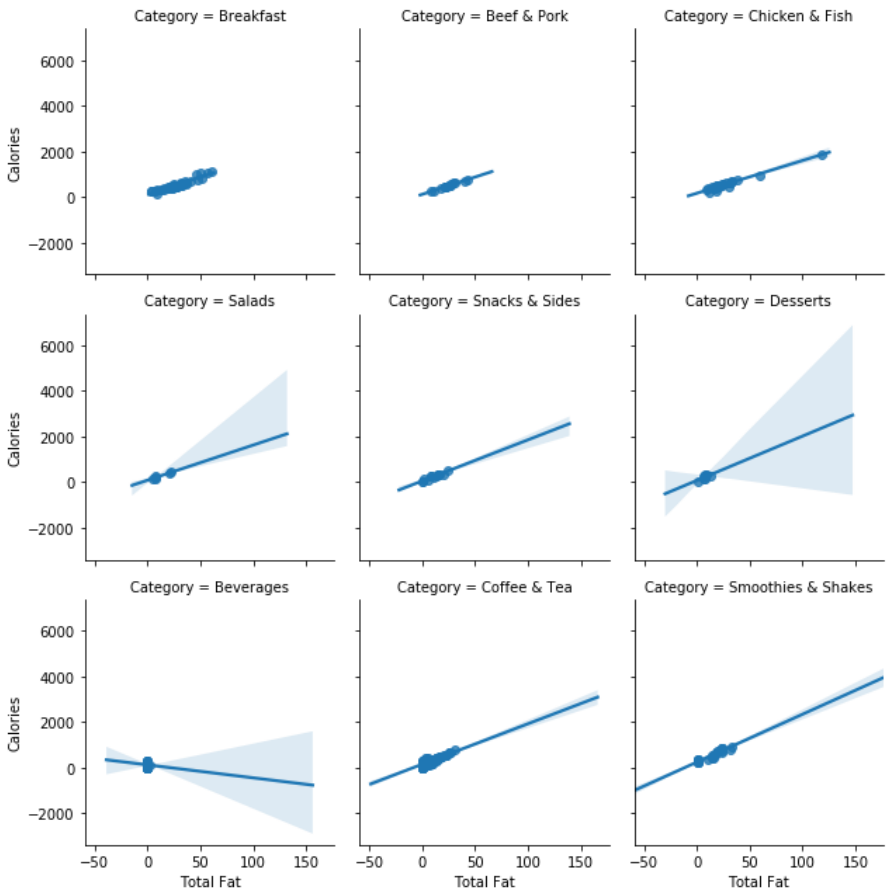
- Cho dữ liệu **menu.csv**, thực hiện các yêu cầu sau :
  1. Đọc dữ liệu, hiển thị thông tin chung của dữ liệu : head, tail, info, describe. (0.25 điểm)
  2. Vẽ biểu đồ phân phối tần suất các món theo Total fat gợi ý như 2 hình sau : (0.5 điểm)



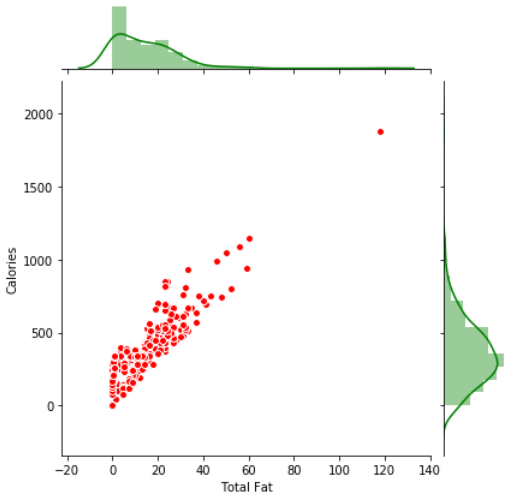
3. Nhận xét biểu đồ trên (0.25 điểm)
4. Tính hệ số tương quan giữa 2 thuộc tính **Total Fat** và **Calories**, sau đó vẽ biểu đồ như hình sau : (0.5 điểm)



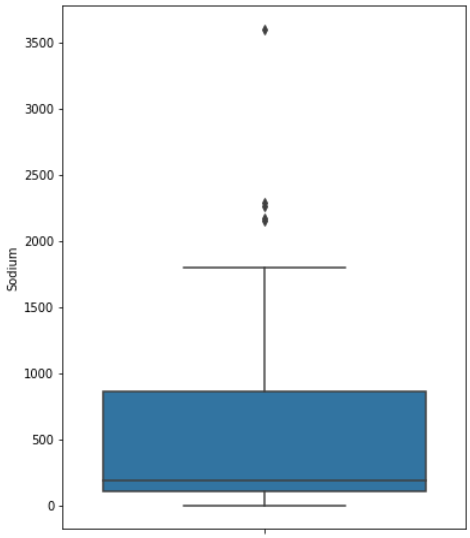
5. Vẽ biểu đồ thể hiện sự tương quan giữa **Total Fat** và **Calories** theo từng nhóm thực phẩm (Category) gợi ý như hình sau : (0.5 điểm)



6. Vẽ biểu đồ thể hiện sự tương quan giữa Total Fat và Calories theo gợi ý như hình sau : (0.5 điểm)



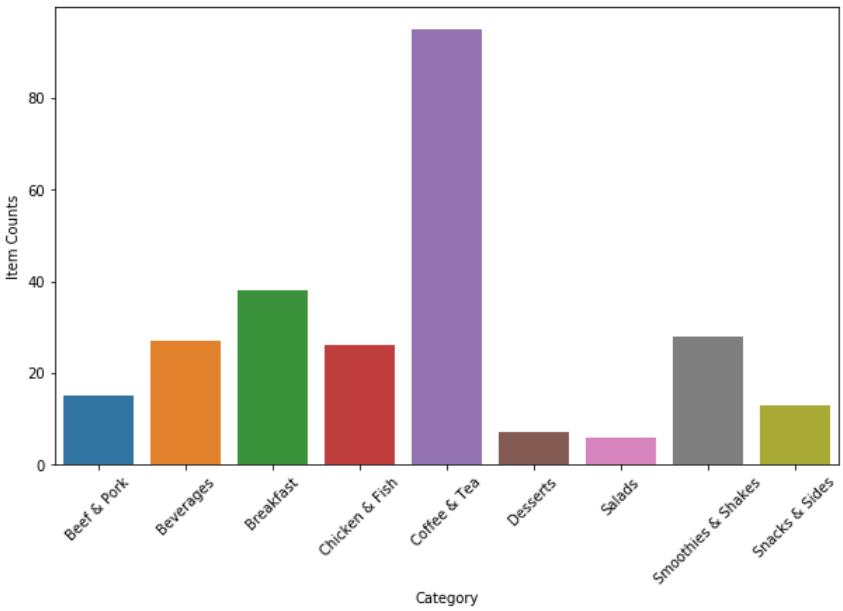
7. Nhận xét biểu đồ trên (0.25 điểm)
8. Vẽ biểu đồ kiểm tra dữ liệu của cột Sodium gợi ý như hình sau : (0.25 điểm)



9. Dữ liệu của cột Sodium theo như hình trên có outliers hay không, nếu có thì loại bỏ tất cả các dòng trong data có outliers? (0.5 điểm)
10. Nhóm dữ liệu Category và đếm theo Item, cho biết mỗi nhóm có bao nhiêu? (0.25 điểm). Gợi ý :

	Category	Item_counts
0	Beef & Pork	15
1	Beverages	27
2	Breakfast	38
3	Chicken & Fish	26
4	Coffee & Tea	95
5	Desserts	7
6	Salads	6
7	Smoothies & Shakes	28
8	Snacks & Sides	13

11. Vẽ biểu đồ thể hiện dữ liệu Category theo gợi ý như hình sau : (0.25 điểm)

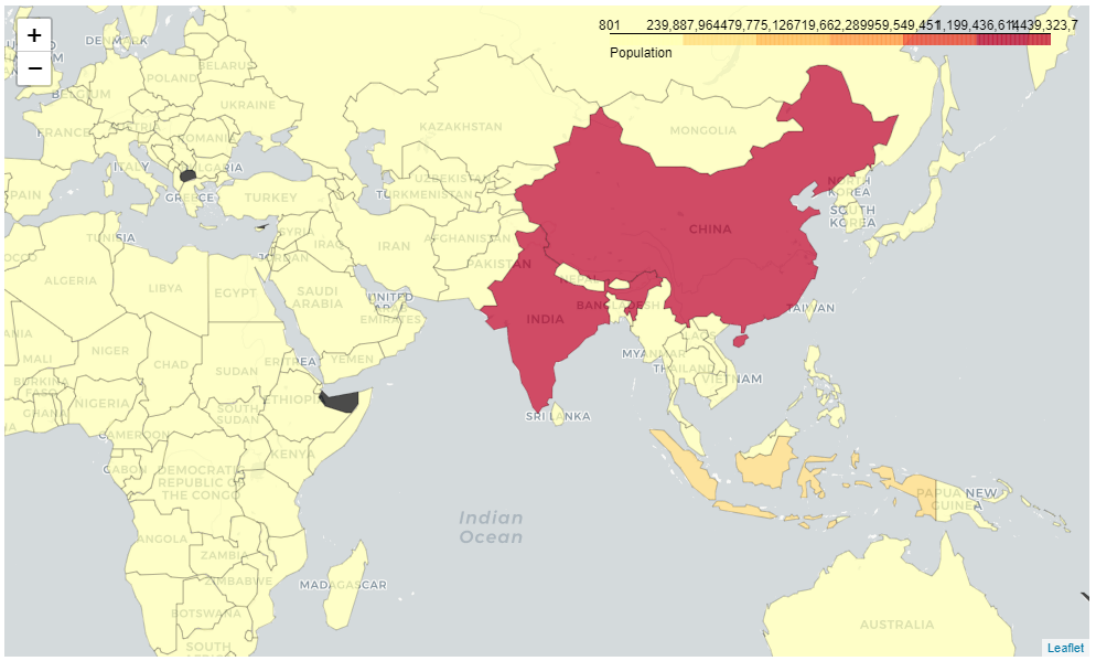


4. Trực quan hóa dữ liệu bản đồ (3 điểm)

- Cho dữ liệu **world_countries_population_2020.csv** và **world-countries.json**, thực hiện các yêu cầu sau :

1. Đọc dữ liệu **world_countries_population_2020.csv**, hiển thị thông tin chung của dữ liệu bao gồm : head, tail, info, describe (0.75 điểm)
2. Chuyển đổi kiểu dữ liệu của cột **Pupulation** và **Lan Area (km2)** sang kiểu số (0.75 điểm)
3. Tạo bản đồ có kiểu **cartodbpositron** với center là Ấn Độ (location=[20.5937, 78.9629]) và zoom level (zoom_start=3) gợi ý như hình sau : (0.75 điểm)

4. Tạo choropleth map theo Population của từng quốc gia theo gợi ý như hình sau : (0.75 điểm)



--- Chúc các bạn làm bài tốt ☺ ---