Đề thi:

PYTHON FOR MACHINE LEARNING, DATA SCIENCE AND VISUALIZATION

Thời gian: 120 phút

Ngày thi : 19/12/2021

**** Học viên tạo 1 thư mục là **LDS2_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)

- Yêu cầu: sử dụng thư viện Numpy thực hiện các yêu cầu sau :
  - Xây dựng hàm kiểm tra số nguyên tố **def kiem_tra_so_nguyen_to (so)** để kiểm tra số truyền vào có phải là số nguyên tố hay không (số nguyên tố là số lớn hơn 1, chỉ chia hết cho 1 và chính nó, ví dụ : 2,3,5,7,11,13….). Kết quả trả về True nếu là số nguyên tố, ngược lại trả về False. (0.5 điểm)
  - Phát sinh mảng 2 chiều có kích thước 4x4 với các phần tử có giá trị phát sinh ngẫu nhiên từ 1 đến 100 (0.5 điểm).
  - Kiểm tra và xuất ra danh sách các phần tử nằm trên đường chéo chính là số nguyên tố (Đường chéo chính chứa các phần tử có chỉ số dòng bằng cột). (0.5 điểm)
- Một số kết quả gợi ý :

```
Danh sách các phần tử được phát sinh ngẫu nhiên trong mảng:
[[ 6 21 13 64]
 [51 33 48 38]
 [67 93 11 35]
 [36 33 34 80]]
Danh sách các phần tử trên đường chéo chính là số nguyên tố
[11]
```

2. Progamming book (1.5 điểm)

- Cho dữ liệu **prog_book.csv** thực hiện các yêu cầu sau :
  a. Đọc dữ liệu và tạo đoạn text từ cột **Book_title**. Sau đó thực hiện chuẩn hóa đoạn text (loại bỏ các từ không quan trọng như of, a ,the,as…) (0.5 điểm)

	Rating	Reviews	Book_title	Description	Number_Of_Pages	Type	Price
0	4.17	3,829	The Elements of Style	This style manual offers practical advice on i...	105	Hardcover	9.323529
1	4.01	1,406	The Information: A History, a Theory, a Flood	James Gleick, the author of the best sellers C...	527	Hardcover	11.000000
2	3.33	0	Responsive Web Design Overview For Beginners	In Responsive Web Design Overview For Beginner...	50	Kindle Edition	11.267647
3	3.97	1,658	Ghost in the Wires: My Adventures as the World...	If they were a hall of fame or shame for compu...	393	Hardcover	12.873529
4	4.06	1,325	How Google Works	Both Eric Schmidt and Jonathan Rosenberg came ...	305	Kindle Edition	13.164706

  b. Tạo biểu đồ Wordcloud có kết quả gợi ý như sau : (0.5 điểm)

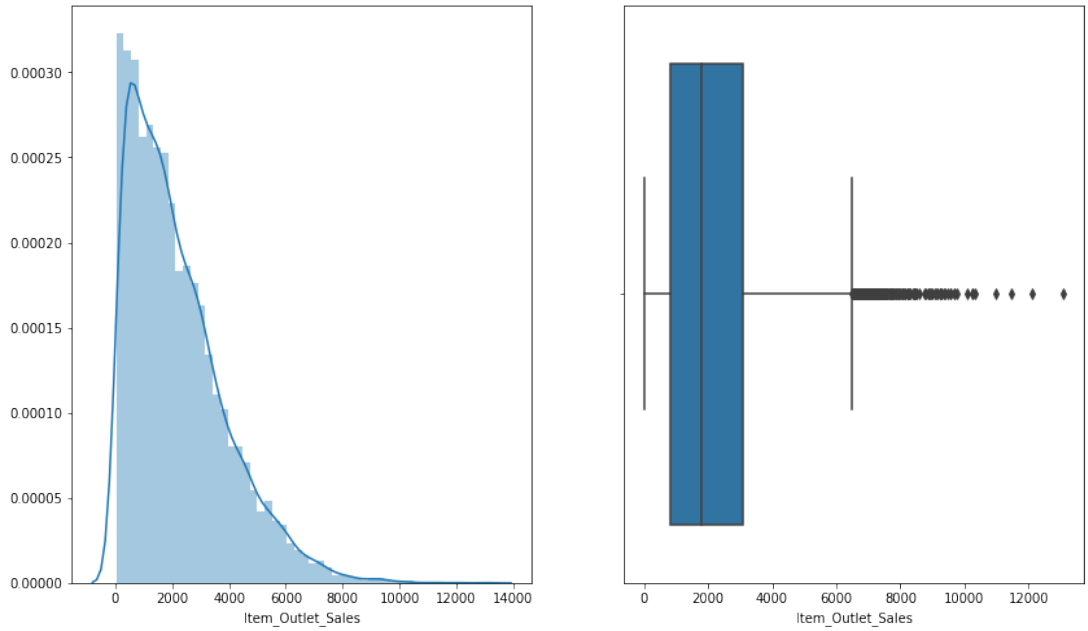
c. Cho tập tin hình ảnh **chiakhoa.jpg**, hãy tạo biểu đồ có kết quả gợi ý như hình sau : (0.5 điểm)

3. Sale : (4 điểm)

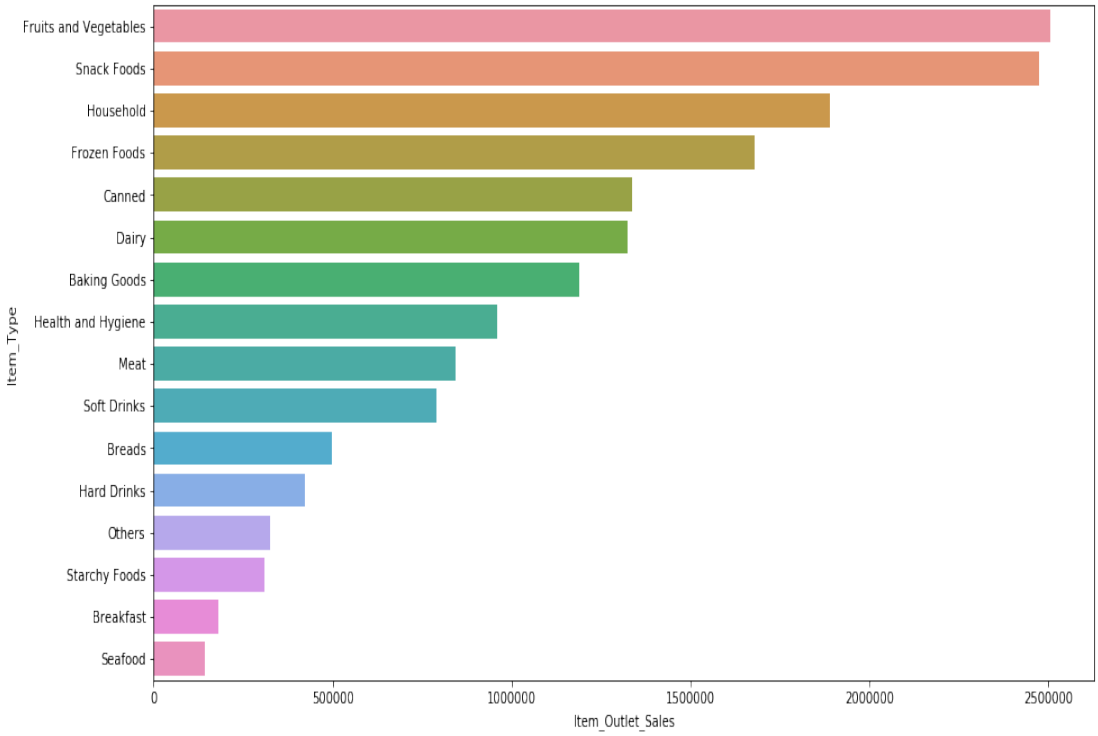
- Cho dữ liệu **sale.csv**, thực hiện các yêu cầu sau :

1. Đọc dữ liệu, hiển thị thông tin chung của dữ liệu : head, tail, info, describe. (0.25 điểm)
2. Kiểm tra xem dữ liệu có bị null ở cột nào hay không ? (0.25 điểm)
3. Thay thế giá trị null ở cột Item_Weight bằng giá trị trung bình của cột này và giá trị null của cột Outlet_Size bằng giá trị xuất hiện nhiều nhất của cột. (0.25 điểm)

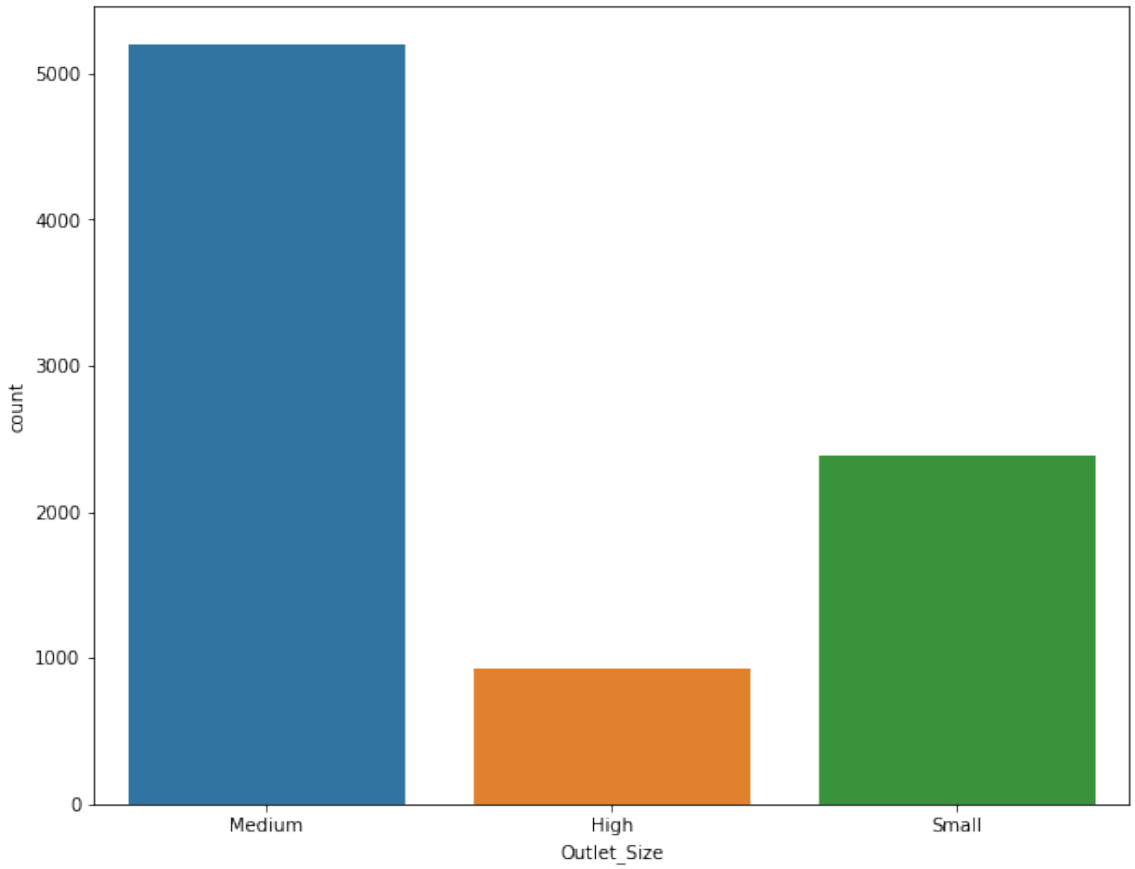
4. Thay thế các giá trị 'low fat' ,'LF' thành 'Low Fat' và 'reg' thành 'Regular' trong cột Item_Fat_Content. (0.25 điểm)
5. Vẽ biểu đồ thể hiện sự phân bố doanh số của các mặt hàng như hình sau và nhận xét. (0.5 điểm)



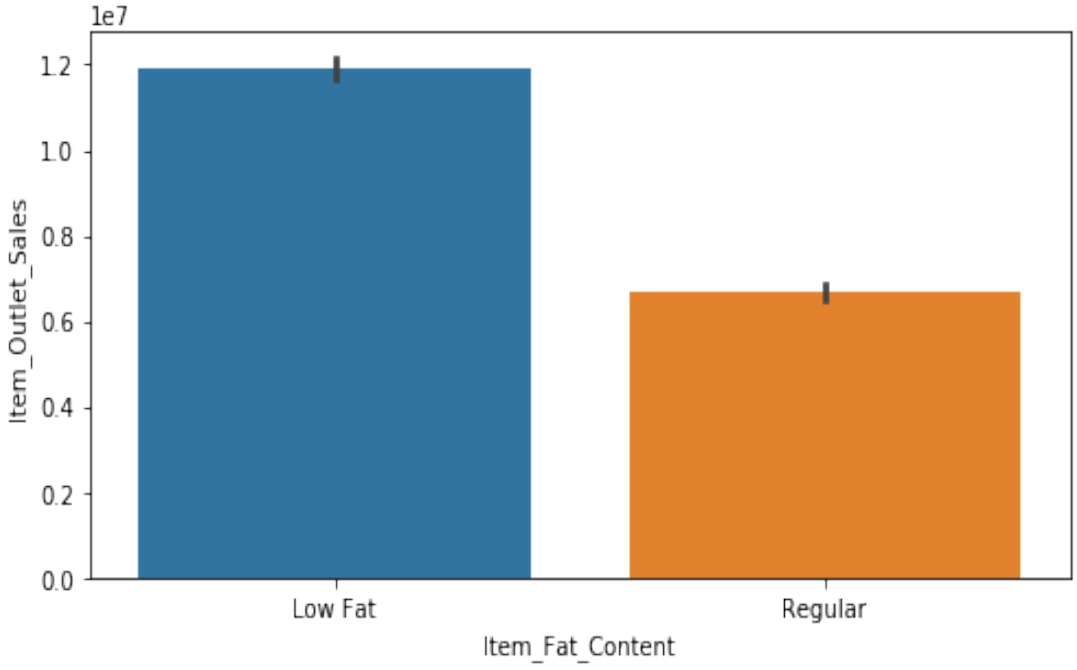
6. Vẽ biểu đồ thể hiện tổng doanh thu của các mặt hàng theo 'Item_Type' và nhận xét. (0.5 điểm)



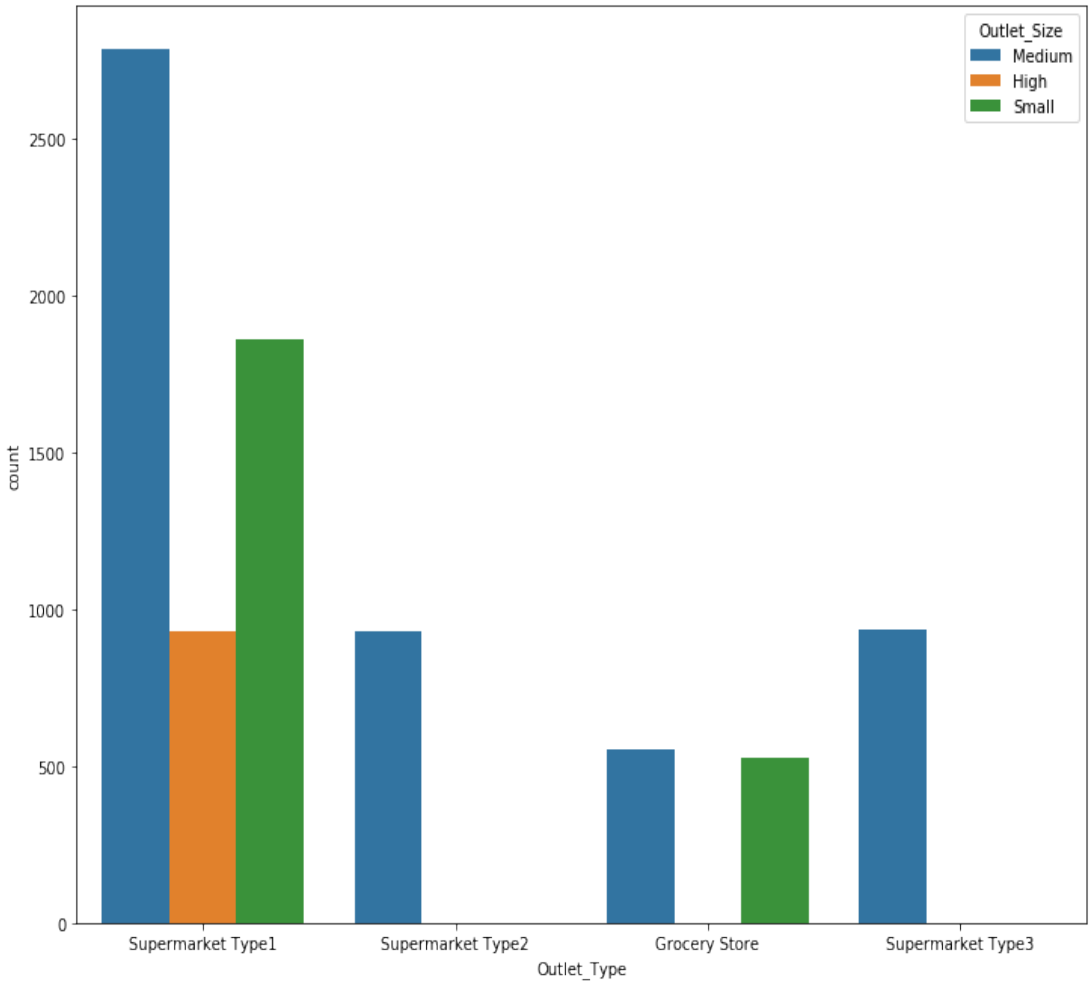
7. Vẽ biểu đồ cho biết số lượng hàng bán được theo từng Outlet_Size và cho biết nhận xét. (0.5 điểm)



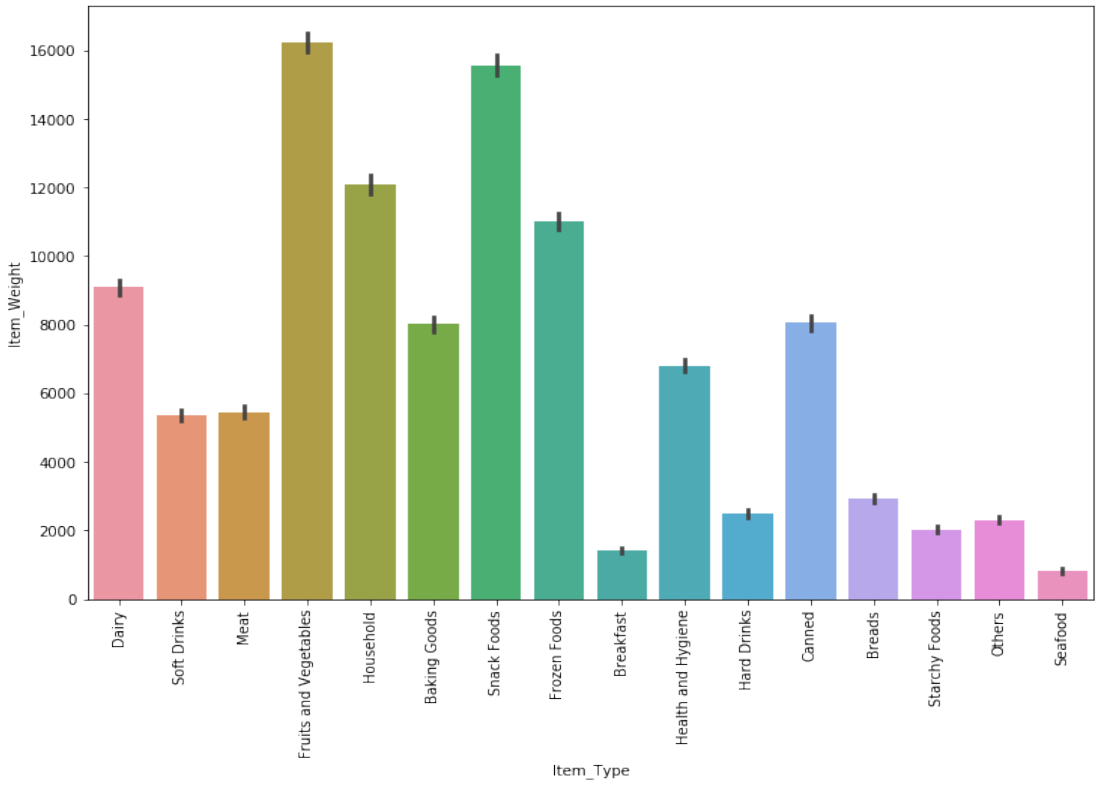
8. Vẽ biểu đồ thể hiện tổng doanh thu của các mặt hàng theo 'Item_Fat_Content' (Gợi ý sử dụng tham số estimator=sum trong barplot của seaborn) (0.25 điểm)



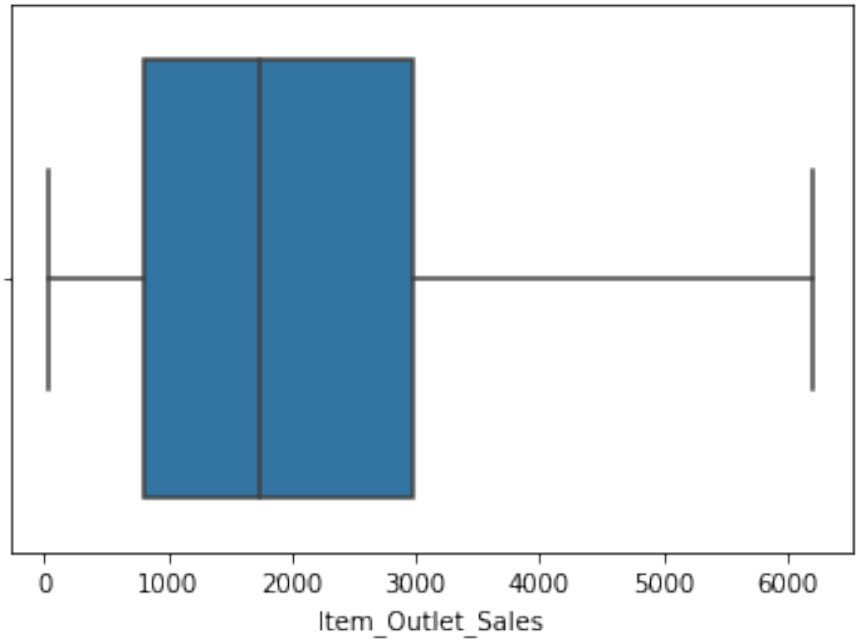
9. Vẽ biểu đếm số loại cửa hàng theo kích cỡ như hình gợi ý. Bạn có nhận xét gì về biểu đồ. (0.25điểm)



10. Vẽ biểu đồ thể hiện tổng trọng lượng của các mặt hàng theo 'Item_Type' (Gợi ý sử dụng tham số estimator=sum trong barplot của seaborn) (0.5 điểm)



11. Dữ liệu của cột 'Item_Outlet_Sales' theo như hình câu 5 có outliers hay không, nếu có thì loại bỏ tất cả các dòng trong data có outliers? (0.5 điểm)



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

- Cho dữ liệu **2014_world_gdp.csv** và **world-countries.json**, thực hiện các yêu cầu sau :

1. Đọc dữ liệu **2014_world_gdp.csv**, hiển thị thông tin chung của dữ liệu bao gồm : head, tail, info, describe (0.75 điểm)

	COUNTRY	GDP_CODE
0	Afghanistan	21.71 AFG
1	Albania	13.4 ALB
2	Algeria	227.8 DZA
3	American Samoa	0.75 ASM
4	Andorra	4.8 AND

2. Thêm cột mới **GDP** được tách ra từ cột **GDP_CODE** (0.5 điểm)

	COUNTRY	GDP_CODE	GDP
0	Afghanistan	21.71 AFG	21.71
1	Albania	13.4 ALB	13.4
2	Algeria	227.8 DZA	227.8
3	American Samoa	0.75 ASM	0.75
4	Andorra	4.8 AND	4.8

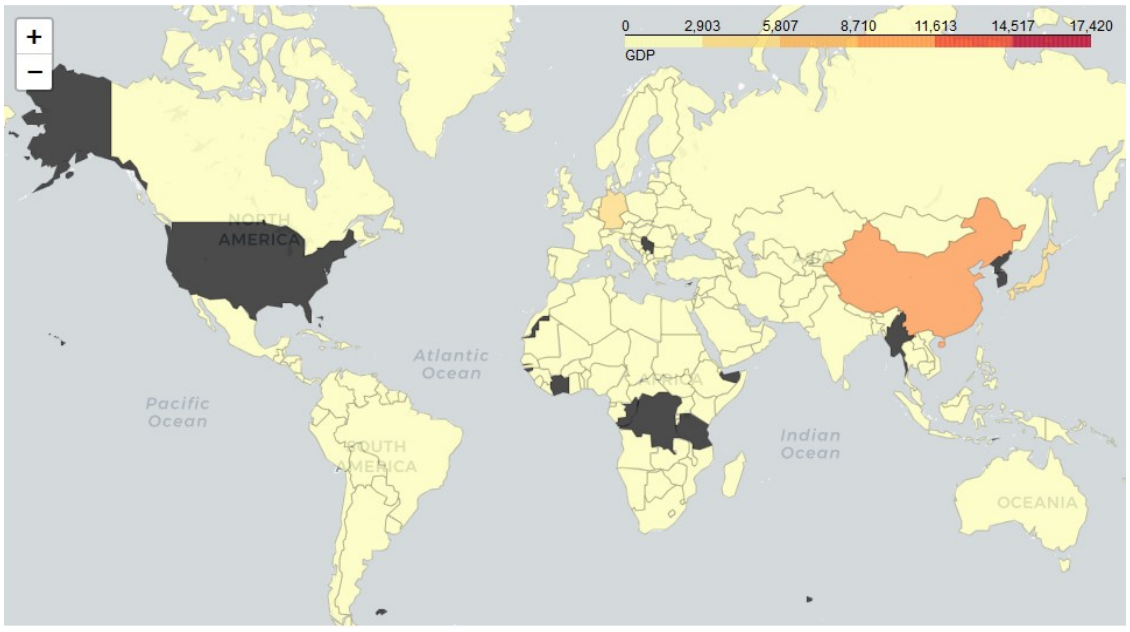
3. Chuyển đổi kiểu dữ liệu của cột **GDP** sang kiểu float64 (0.25 điểm)

```
1  df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 222 entries, 0 to 221
Data columns (total 3 columns):
COUNTRY     222 non-null object
GDP_CODE    222 non-null object
GDP         222 non-null float64
dtypes: float64(1), object(2)
memory usage: 5.2+ KB
```

4. Tạo bản đồ có kiểu **cartodbpositron** với center (location=[0, 0]) và zoom level (zoom_start=3) gợi ý như hình sau : (0.75 điểm)

5. Tạo choropleth map theo **GDP** của từng quốc gia theo gợi ý như hình sau : (0.75 điểm)



--- Chúc các bạn làm bài tốt ☺ ---