Đề thi:

PYTHON FOR MACHINE LEARNING, DATA SCIENCE AND VISUALIZATION

Thời gian: 120 phút

Ngày thi : 11/12/2022

**** 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.

Câu 1. Numpy Array (1.5 điểm)

1. Yêu cầu: sử dụng thư viện Numpy thực hiện các yêu cầu sau:

– Tạo mảng arr1 ngẫu nhiên 10 phần tử số nguyên từ 1-20 với np.random.seed(3)

(0.25 điểm)

array([20, 18, 19, 13, 18, 18, 10, 15, 13, 20])

– Phát sinh mảng arr2 ngẫu nhiên 10 phần tử số nguyên từ 11-30 với np.random.seed(3) (0. 25 điểm)

array([21, 14, 19, 11, 30, 21, 22, 20, 21, 17])

– In ra các phần tử xuất hiện trong cả 2 mảng arr1, arr2 (0.25 điểm)

array([19, 20])

– Sắp xếp lại mảng arr1 theo thứ tự tăng dần (0.25 điểm)

array([10, 13, 13, 15, 18, 18, 18, 19, 20, 20])

– In ra các phần tử trong mảng arr2 là số nguyên tố (0.5 điểm)

array([19, 11, 17])

Câu 2: Text (tech.csv)(1.5 điểm)

Cho dữ liệu **tech.csv** thực hiện các yêu cầu sau :
1. Đọc dữ liệu và tạo đoạn text từ cột content. Sau đó thực hiện chuẩn hóa đoạn text (loại bỏ các từ không quan trọng) (0.5 điểm)
Gợi ý: dùng stopwords và bổ sung thêm 'The', 'people', 'U', 'will', 'one', 'much', 'many',...

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

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

Câu 3. Xử lý dữ liệu và trực quan hóa: (5 điểm)

1. 1.Tạo các DataFrame products, sales, stores chứa thông tin các sản phẩm, thông tin bán hàng và thông tin cửa hàng. Các danh sách này được đọc từ các tập tin products.csv, sales.csv, stores.csv (0.5 điểm)

2. Hiển thị 5 dòng đầu và thông tin info của mỗi dataframe (0.5 điểm)

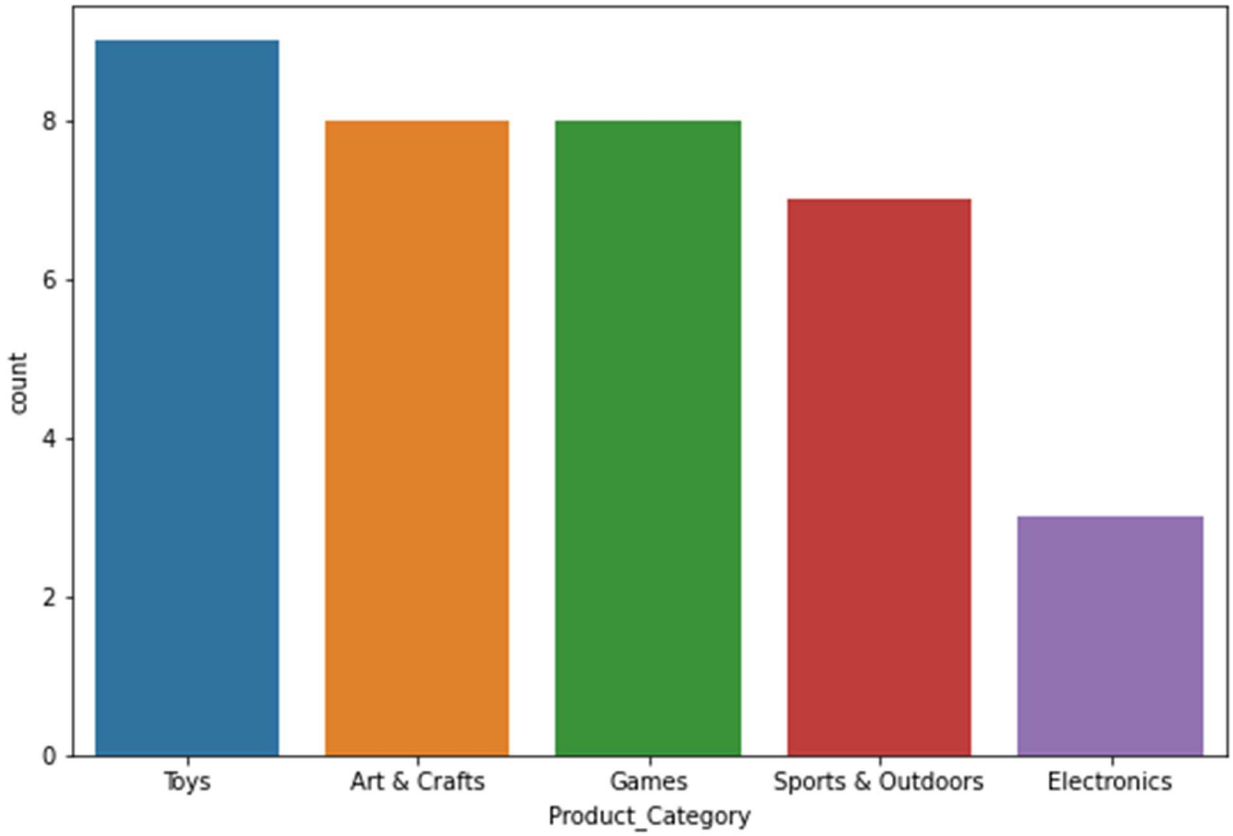
	Product_ID	Product_Name	Product_Category	Product_Cost	Product_Price
0	1	Action Figure	Toys	$9.99	$15.99
1	2	Animal Figures	Toys	$9.99	$12.99
2	3	Barrel O' Slime	Art & Crafts	$1.99	$3.99
3	4	Chutes & Ladders	Games	$9.99	$12.99
4	5	Classic Dominoes	Games	$7.99	$9.99

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 35 entries, 0 to 34
Data columns (total 5 columns):
 #   Column            Non-Null Count  Dtype 
---  ------            --------------  ----- 
 0   Product_ID        35 non-null     int64 
 1   Product_Name      35 non-null     object
 2   Product_Category  35 non-null     object
 3   Product_Cost      35 non-null     object
 4   Product_Price     35 non-null     object
dtypes: int64(1), object(4)
memory usage: 904.0+ bytes
```

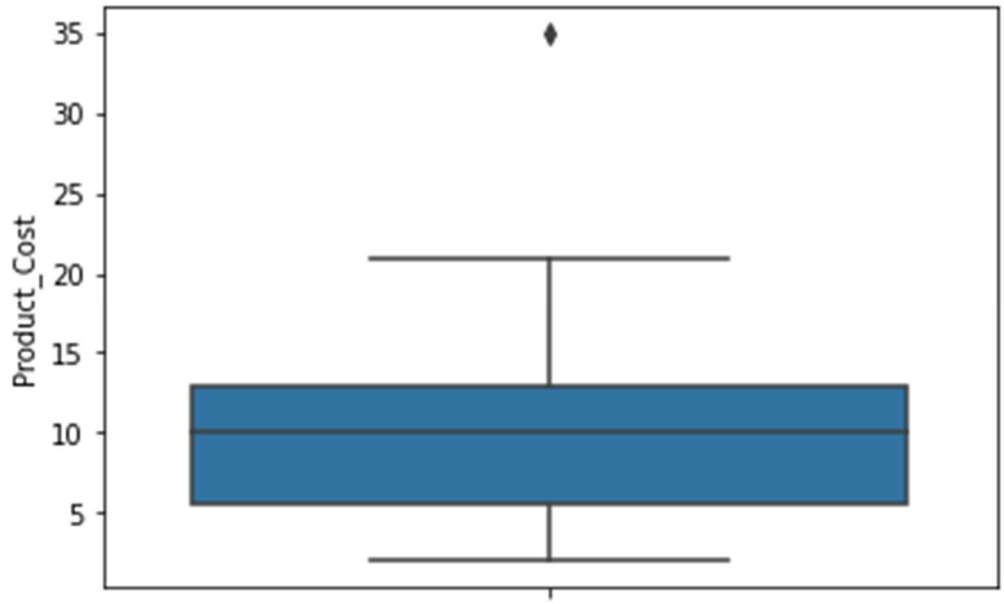
3. Từ df products, chuyển cột 'Product_Price', 'Product_Cost' sang kiểu float (0.25 điểm)

	Product_ID	Product_Name	Product_Category	Product_Cost	Product_Price
0	1	Action Figure	Toys	9.99	15.99
1	2	Animal Figures	Toys	9.99	12.99
2	3	Barrel O' Slime	Art & Crafts	1.99	3.99
3	4	Chutes & Ladders	Games	9.99	12.99
4	5	Classic Dominoes	Games	7.99	9.99

4. Vẽ biểu đồ thể hiện số lượng sản phẩm theo Product_Category. Nhận xét (0.25 điểm)

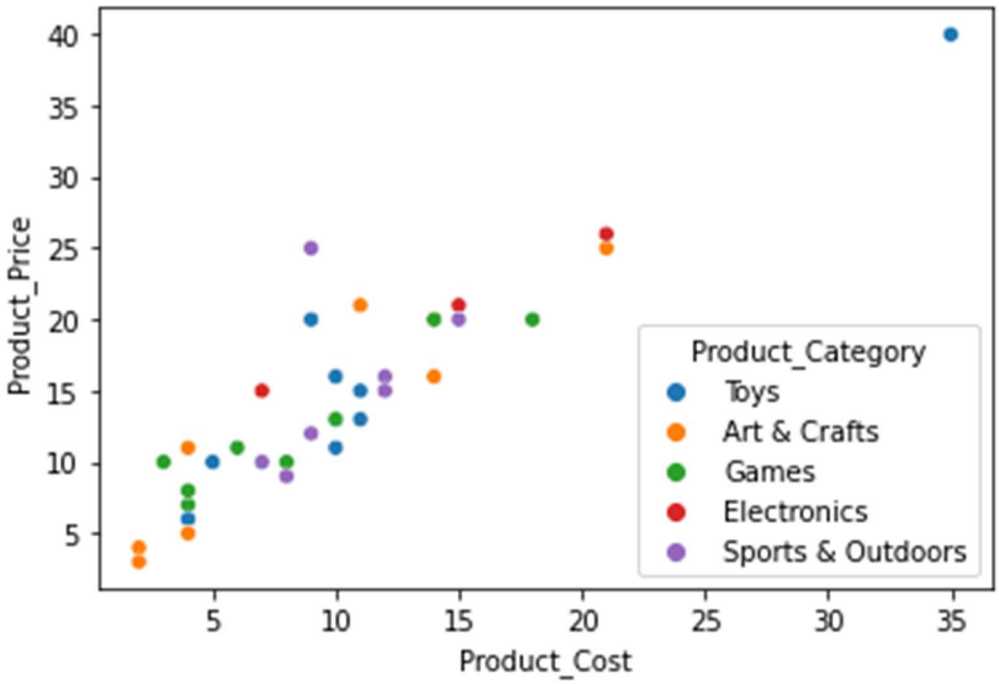


5. Vẽ biểu đồ boxplot với dữ liệu là Product_Cost. Chép các outlier ra df_outliers (0.25 điểm)



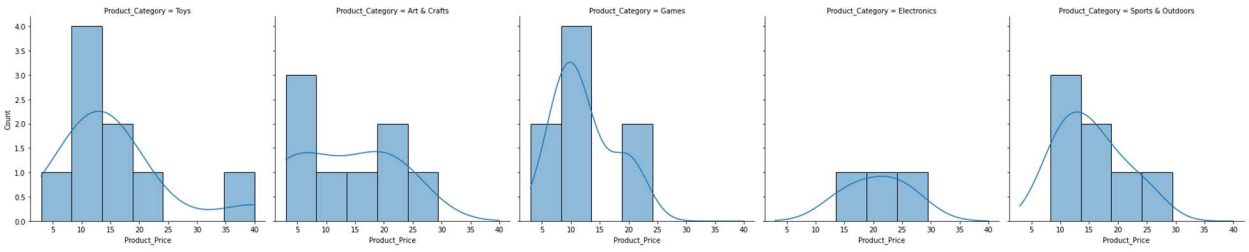
	Product_ID	Product_Name	Product_Category	Product_Cost	Product_Price
17	18	Lego Bricks	Toys	34.99	39.99

6. Vẽ biểu đồ scatter plot thể hiện mối tương quan giữa chi phí sản xuất và giá bán sản phẩm. Cho biết hệ số tương quan và nhận xét (0.5điểm)



	Product_Cost	Product_Price
Product_Cost	1.000000	0.906848
Product_Price	0.906848	1.000000

7. Vẽ biểu đồ phân bố giá bán của mỗi nhóm hàng. Nhận xét (0.25 điểm)



8. Tạo dataframe df bằng cách trộn products, sales, stores theo how = 'inner'. Chỉ lấy các cột Product_Category, Product_Cost, Product_Price, Sale_ID, Date, Store_ID, Units. Hiển thị head, info: (0.25 điểm)

	Product_Category	Product_Cost	Product_Price	Sale_ID	Date	Store_ID	Units
0	Toys	9.99	15.99	2	2017-01-01	28	1
1	Toys	9.99	15.99	34	2017-01-01	36	1
2	Toys	9.99	15.99	47	2017-01-01	30	3
3	Toys	9.99	15.99	59	2017-01-01	41	1
4	Toys	9.99	15.99	61	2017-01-01	36	1

9. Tạo cột 'Sales' = 'Units' * 'Product_Price'

'Cost' = 'Units' * 'Product_Cost'

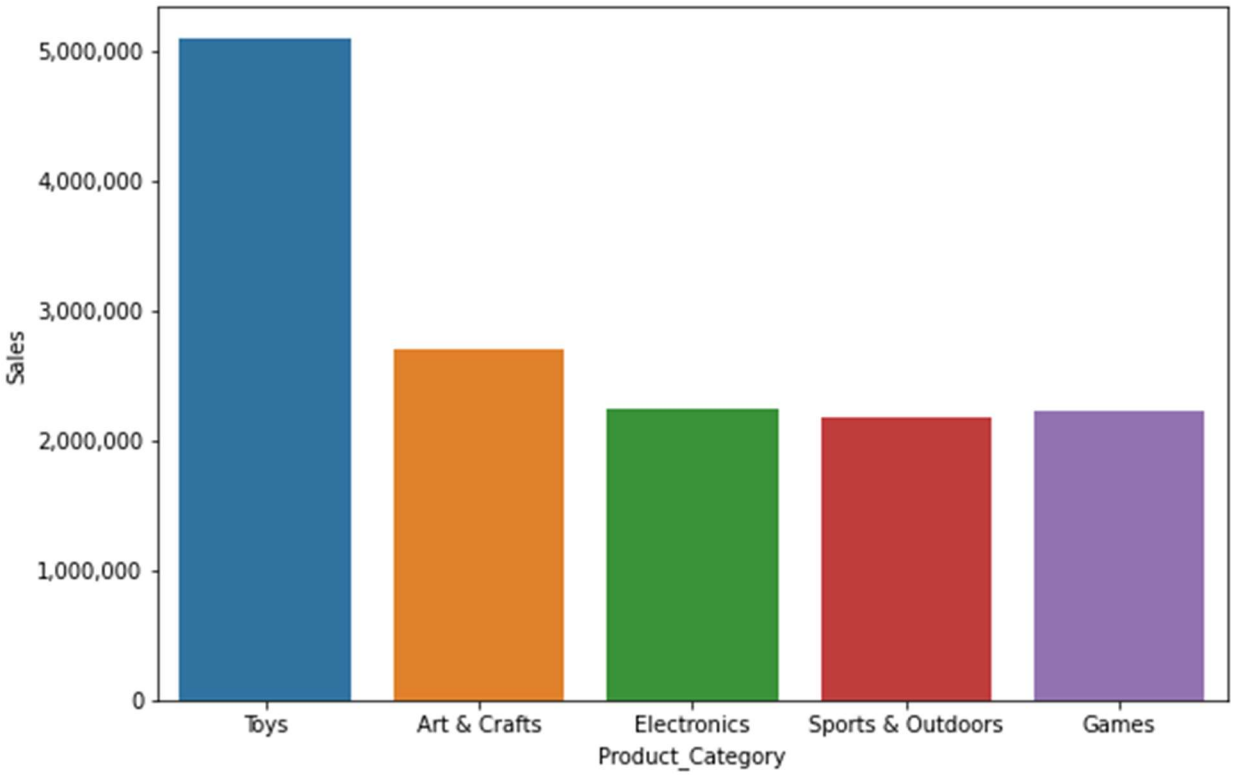
'Profit' = 'Sales' - Cost' (0.5 điểm)

	Product_ID	Product_Name	Product_Category	Product_Cost	Product_Price	Sale_ID	Date	Store_ID	Units	Store_Name	Sales	Cost	Profit
0	1	Action Figure	Toys	9.99	15.99	2	2017-01-01	28	1	Maven Toys Puebla 2	15.99	9.99	6.0
1	1	Action Figure	Toys	9.99	15.99	171	2017-01-01	28	1	Maven Toys Puebla 2	15.99	9.99	6.0
2	1	Action Figure	Toys	9.99	15.99	364	2017-01-01	28	1	Maven Toys Puebla 2	15.99	9.99	6.0
3	1	Action Figure	Toys	9.99	15.99	437	2017-01-01	28	1	Maven Toys Puebla 2	15.99	9.99	6.0
4	1	Action Figure	Toys	9.99	15.99	686	2017-01-01	28	1	Maven Toys Puebla 2	15.99	9.99	6.0

10. Tạo bảng tính tổng doanh số theo từng cửa hàng và nhóm hàng (0.5 điểm)

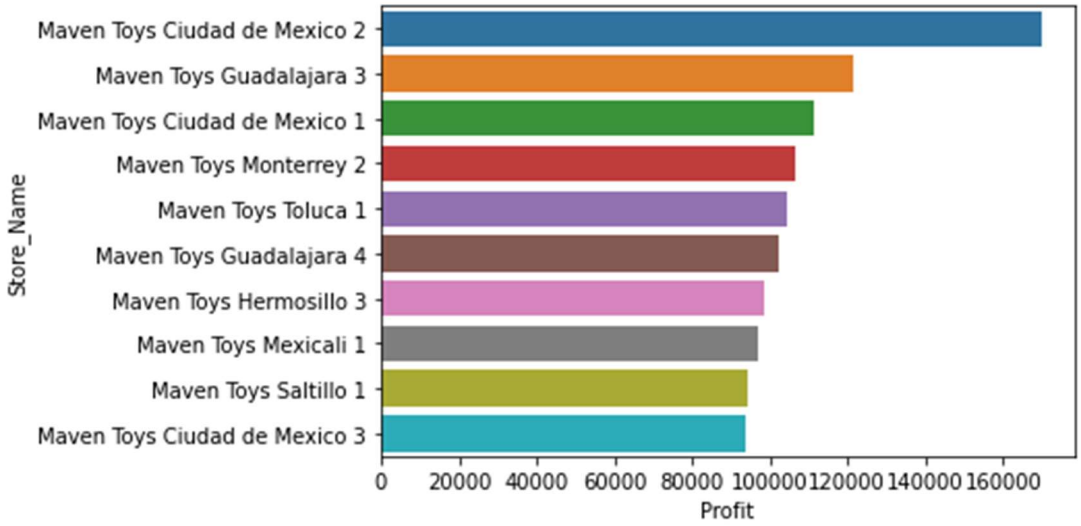
Product_Category Store_ID	Art & Crafts	Electronics	Games	Sports & Outdoors	Toys	All
1	60242	22958	47414	40246	90980	261842
2	52697	43667	32749	34403	114440	277959
3	52375	49580	40552	32838	87088	262435
4	50556	61207	55236	48674	114733	330408
5	41824	26575	34230	33251	75016	210897
6	48109	104047	45010	35691	61159	294019
7	64691	67519	61643	45251	133893	372998
8	60647	24851	39257	45394	67525	237676
9	77326	42529	72040	72838	168819	433556
10	52691	26754	62632	80893	88814	311786

11. Vẽ biểu đồ barplot thể hiện tổng doanh số theo nhóm hàng. Nhận xét (0.25 điểm)



12. Tính và vẽ biểu đồ thể hiện 10 cửa hàng có lợi nhuận cao nhất. Store_Name được lấy từ df stores (0.5 điểm)

	Store_ID	Profit	Store_Name
0	31	169856.0	Maven Toys Ciudad de Mexico 2
1	30	121571.0	Maven Toys Guadalajara 3
2	9	111296.0	Maven Toys Ciudad de Mexico 1
3	7	106783.0	Maven Toys Monterrey 2
4	17	104612.0	Maven Toys Toluca 1
5	46	102178.0	Maven Toys Guadalajara 4
6	42	98825.0	Maven Toys Hermosillo 3
7	6	97206.0	Maven Toys Mexicali 1
8	4	94252.0	Maven Toys Saltillo 1
9	37	94021.0	Maven Toys Ciudad de Mexico 3



13. Vẽ biểu đồ treemap thể hiện đóng góp của 10 cửa hàng có lợi nhuận cao nhất: (0.25 điểm)

14. Vẽ biểu đồ lineplot thể hiện biến động doanh số của 5 nhóm hàng. Nhận xét: (0.25 điểm)

Câu 4: Trực quan hóa dữ liệu bản đồ (2 điểm)

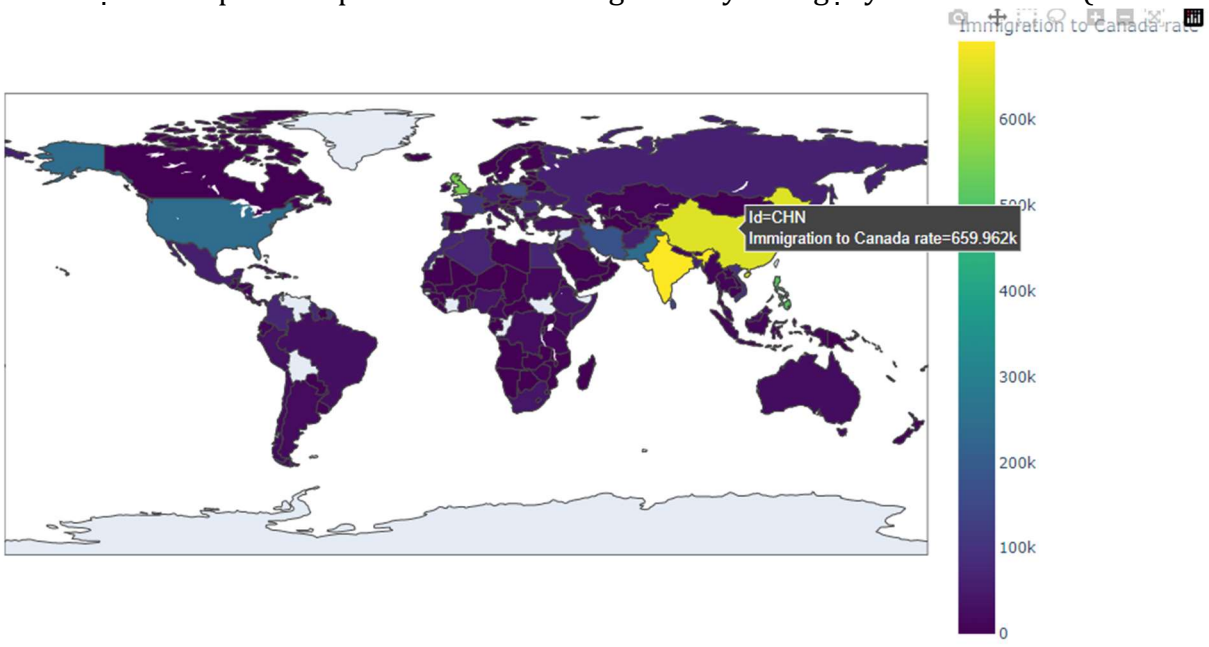
1. Đọc dữ liệu canada.csv, hiển thị thông tin chung của dữ liệu bao gồm: head, tail, info, describe (0.5 điểm)

	Country	Continent	Region	DevName	Total
0	Afghanistan	Asia	Southern Asia	Developing regions	58639.0
1	Albania	Europe	Southern Europe	Developed regions	15699.0
2	Algeria	Africa	Northern Africa	Developing regions	69439.0
3	American Samoa	Oceania	Polynesia	Developing regions	6.0
4	Andorra	Europe	Southern Europe	Developed regions	15.0

2. Đọc tập tin world-countries.json. Tạo cột Id tương ứng với tên Country cho dataframe (0.5 điểm)

	Country	Continent	Region	DevName	Total	Id
0	Afghanistan	Asia	Southern Asia	Developing regions	58639.0	AFG
1	Albania	Europe	Southern Europe	Developed regions	15699.0	ALB
2	Algeria	Africa	Northern Africa	Developing regions	69439.0	DZA
3	American Samoa	Oceania	Polynesia	Developing regions	6.0	None
4	Andorra	Europe	Southern Europe	Developed regions	15.0	None

3. Tạo Choropleth map theo Total của từng country theo gợi ý như hình sau: (1.0 điểm)



--- Chúc các bạn làm bài tốt ☺---