Đề thi:

PYTHON FOR MACHINE LEARNING, DATA SCIENCE AND VISUALIZATION

Thời gian: 120 phút

Ngày thi : 21/08/2022

**** 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.

1. Numpy Array (1.5 điểm)

1. Yêu cầu: sử dụng thư viện Numpy thực hiện các yêu cầu sau :
   - Phát sinh mảng 2 chiều có kích thước 4x4 với các phần tử có giá trị phát sinh ngẫu nhiên từ 1 đến 100 với np.random.seed(1). (0.75 điểm)
   - Thay thế tất cả các giá trị trong mảng để đảm bảo các giá trị phần tử trong mảng thuộc khoảng [10,50]. Thực hiện bằng cách thay tất cả các giá trị lớn hơn 50 thành 50 và dưới 10 thành 10, các giá trị còn lại giữ nguyên. (0.75 điểm)
2. Một số kết quả gợi ý :

   Danh sách các phần tử được phát sinh ngẫu nhiên trong mảng:
   [[38 13 73 10]
   [76 6 80 65]
   [17 2 77 72]
   [7 26 51 21]]

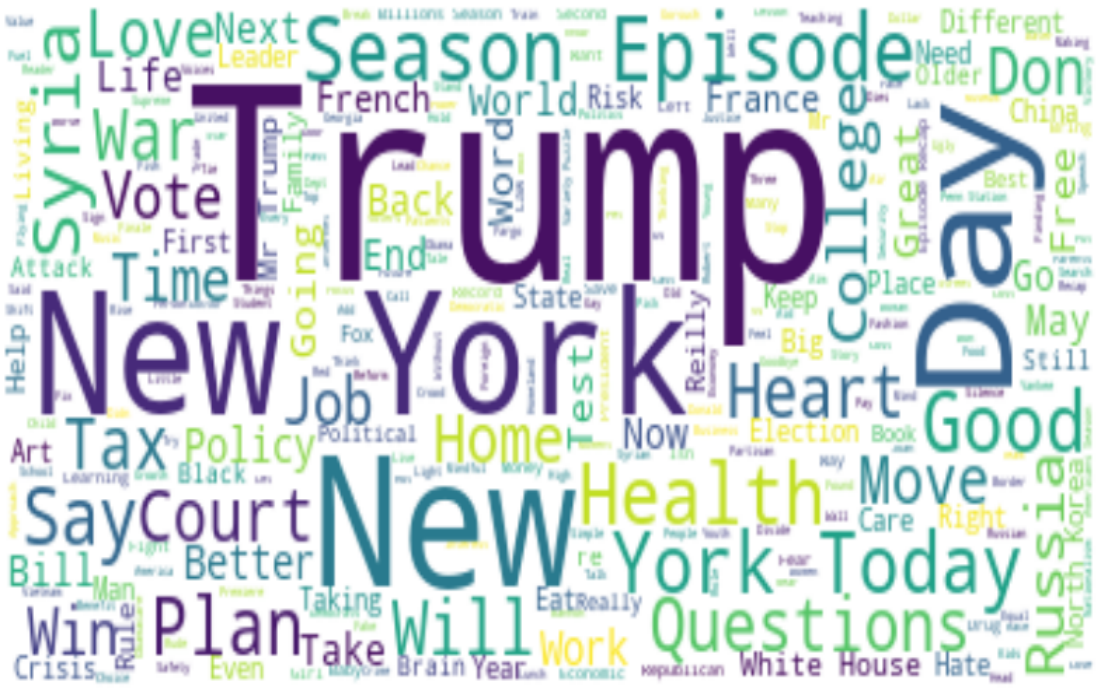
   Mảng sau khi thay thế các giá trị để đảm bảo các phần tử trong mảng thuộc khoảng [10,50] :
   [[38 13 50 10]
   [50 10 50 50]
   [17 10 50 50]
   [10 26 50 21]]

2. New York time comments (1.5 điểm)

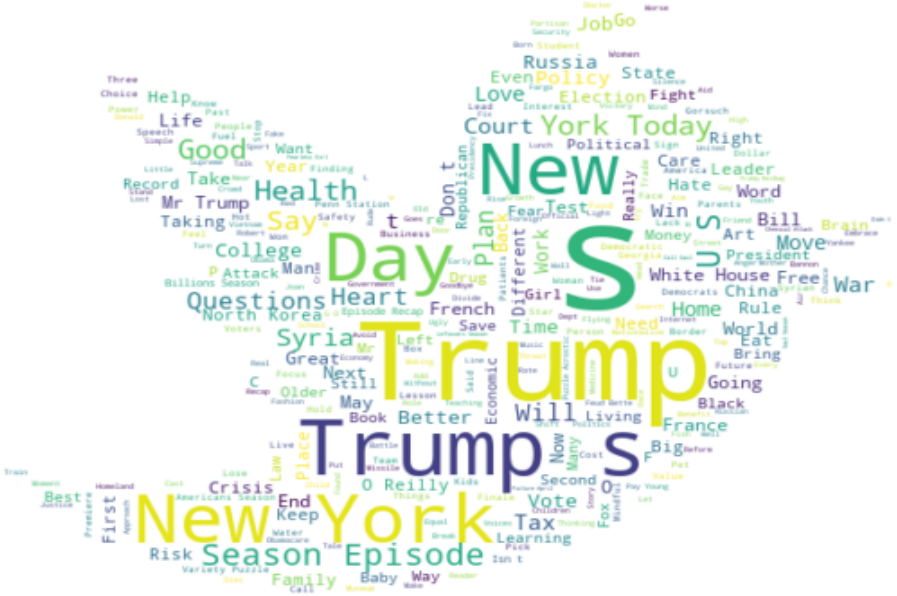
Cho dữ liệu **ArticlesApril2017.csv** thực hiện các yêu cầu sau :

1. Đọc dữ liệu và tạo đoạn text từ cột headline. Sau đó thực hiện chuẩn hóa đoạn text (loại bỏ các từ không quan trọng) (0.5 điểm)

   Vd : Các từ không quan trọng ['one', 'br', 'Po', 'th', 'sayi', 'fo', 'Unknown']
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)



3. **Xử lý dữ liệu và trực quan hóa : (4 điểm)**

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

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

2. Thay thế các giá trị trên cột experience_level với các giá trị sau : (0.25 điểm)

   'EN':'Entry Level/Junior',

   'MI':'Mid Level/Intermediate',

   'SE':'Senior Level/Expert'

   'EX':'Executive Level/Director'

3. Thay thế các giá trị trên cột employment_type với các giá trị sau : (0.25 điểm)

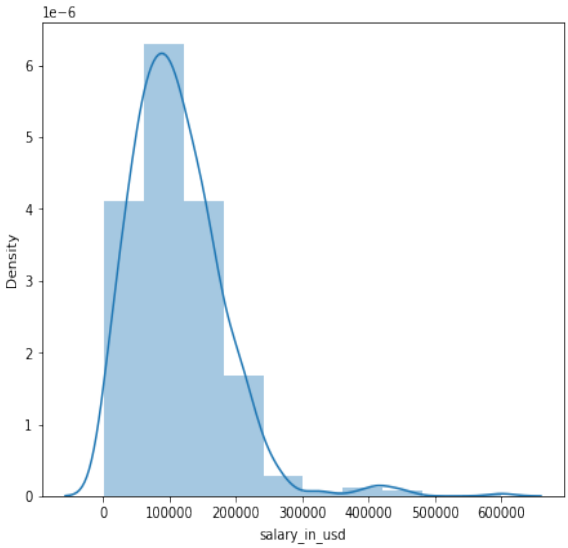
   'FT':'Full Time',

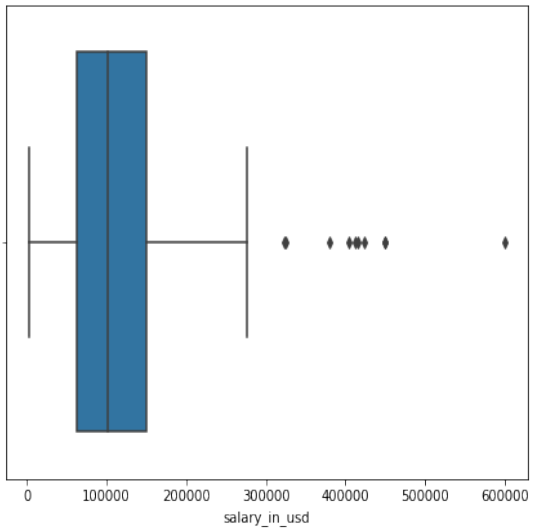
   'PT':'Part Time',

   'CT':'Contract',

   'FL':'Freelance'

4. Vẽ biểu đồ thể hiện sự phân bố lương theo gợi ý như hình sau và nhận xét. (0.5 điểm)

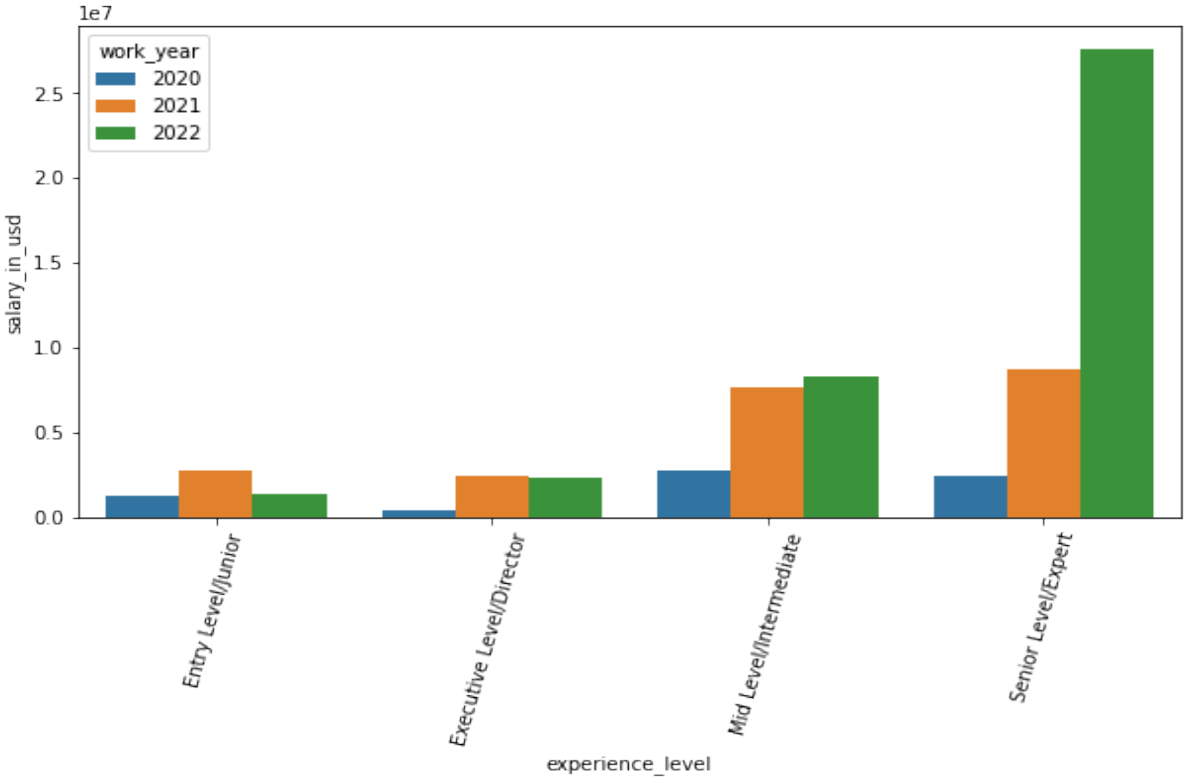




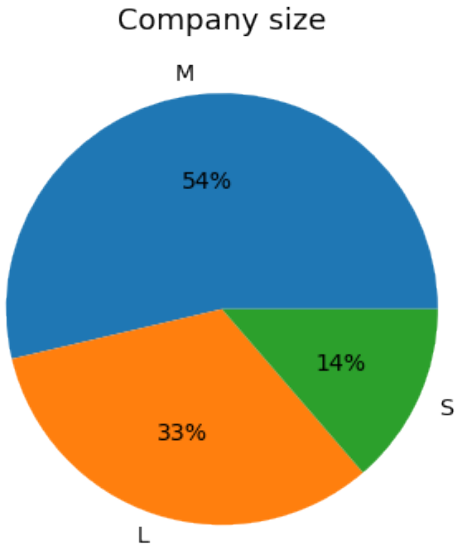
5. Cho biết tổng lương qua các năm theo từng cấp độ của kinh nghiệm công việc. (0.5 điểm)

	experience_level	work_year	salary_in_usd
0	Entry Level/Junior	2020	1272972
1	Entry Level/Junior	2021	2777748
2	Entry Level/Junior	2022	1373892
3	Executive Level/Director	2020	404833
4	Executive Level/Director	2021	2461280
5	Executive Level/Director	2022	2318080
6	Mid Level/Intermediate	2020	2750402
7	Mid Level/Intermediate	2021	7694108
8	Mid Level/Intermediate	2022	8298650
9	Senior Level/Expert	2020	2470329
10	Senior Level/Expert	2021	8735137
11	Senior Level/Expert	2022	27607376

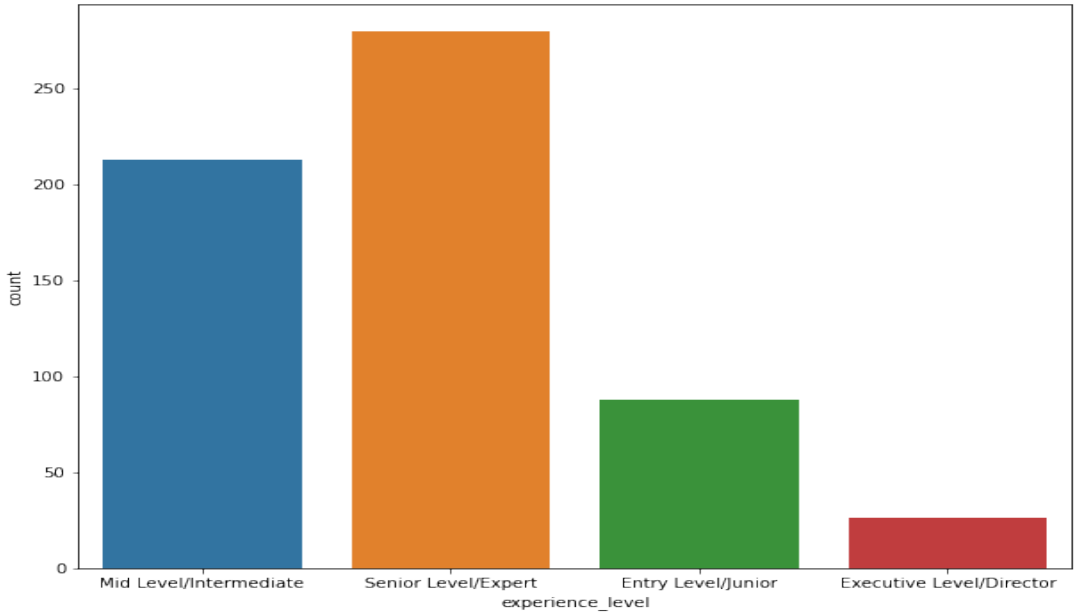
6. Dựa vào dữ liệu câu 5 hãy vẽ biểu đồ thể hiện tổng lương qua các năm theo từng cấp độ kinh nghiệm như gợi ý và nhận xét : (0.25 điểm)



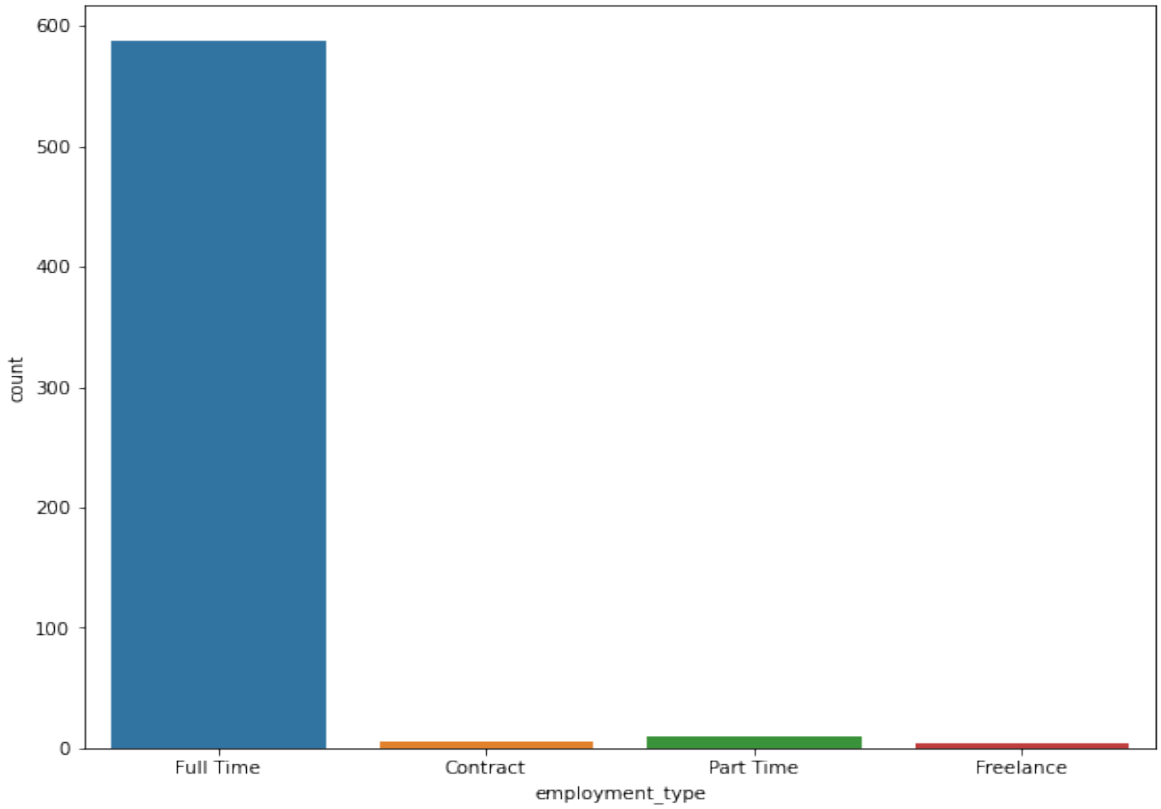
7. Vẽ biểu đồ thể hiện tỉ lệ quy mô các công ty tuyển dụng. (0.25 điểm)



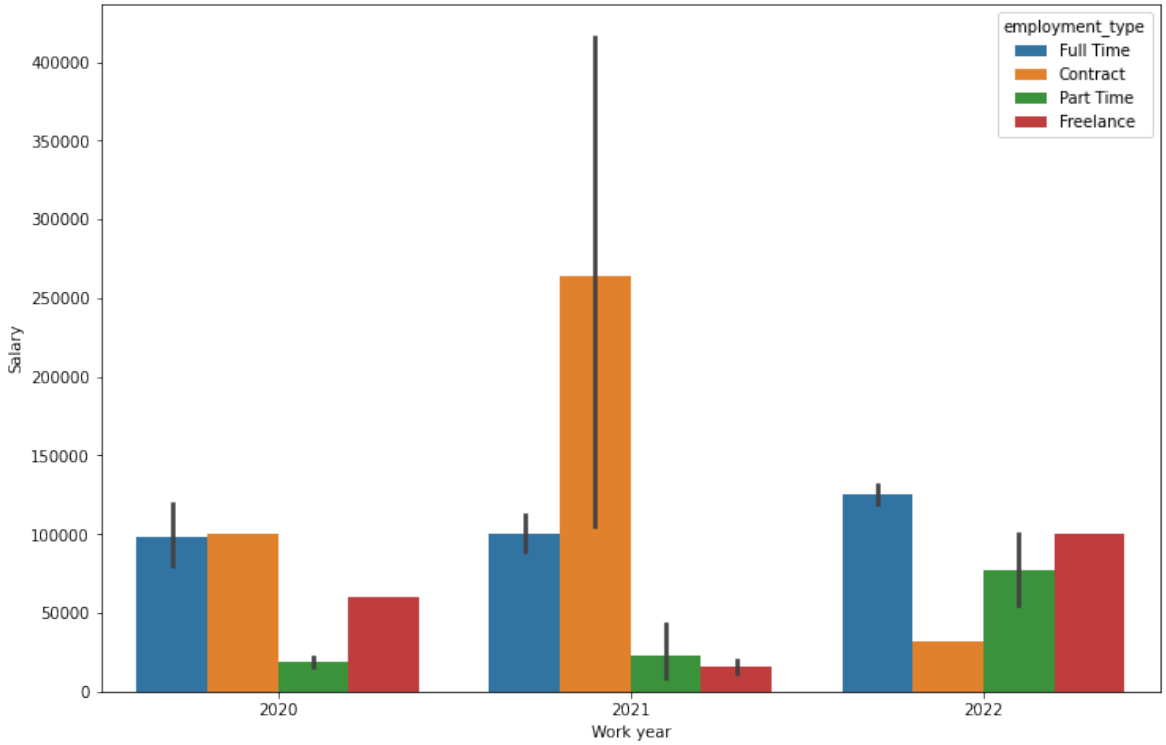
8. Vẽ biểu đồ cho biết số lượng các công việc ứng với các mức kinh nghiệm làm việc. Bạn có nhận xét gì về biểu đồ này. (0.25 điểm)



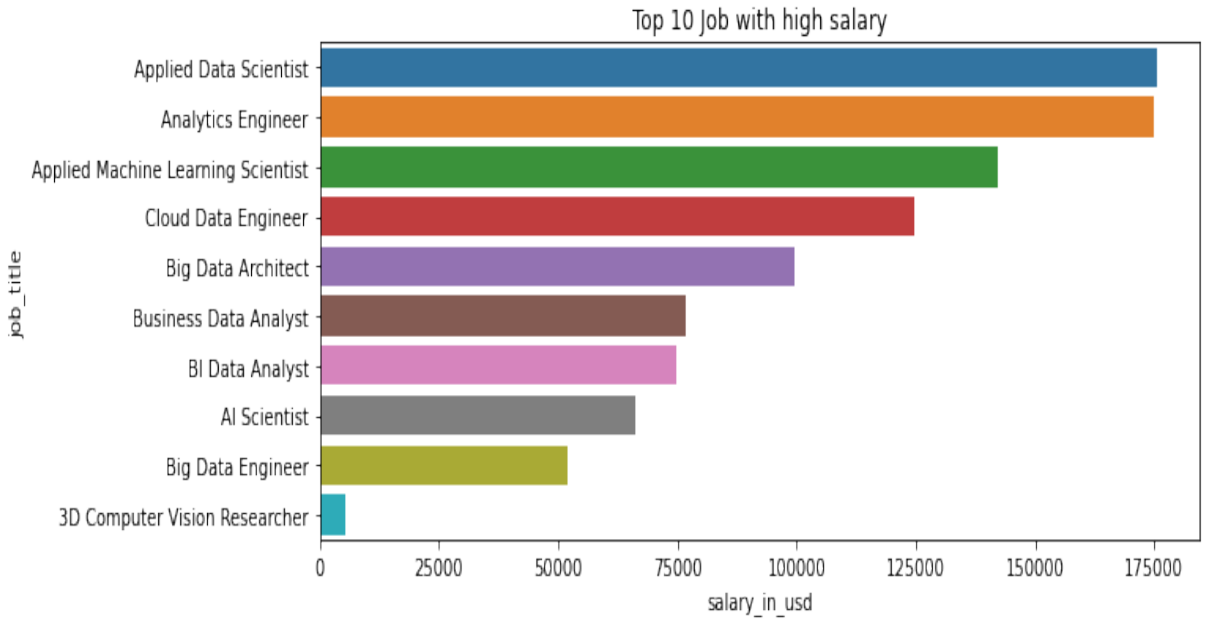
9. Vẽ biểu đồ cho biết số lượng các công việc ứng với các hình thức công việc.. Sau đó nhận xét về biểu đồ: (0.25 điểm)



10. Vẽ biểu đồ cho biết lương trung bình qua các năm ứng với từng hình thức công việc gợi ý như hình sau : (0.5 điểm)



11. Vẽ biểu đồ cho biết 10 công việc có lương trung bình cao nhất gợi ý như hình sau và nhận xét : (0.5 điểm)

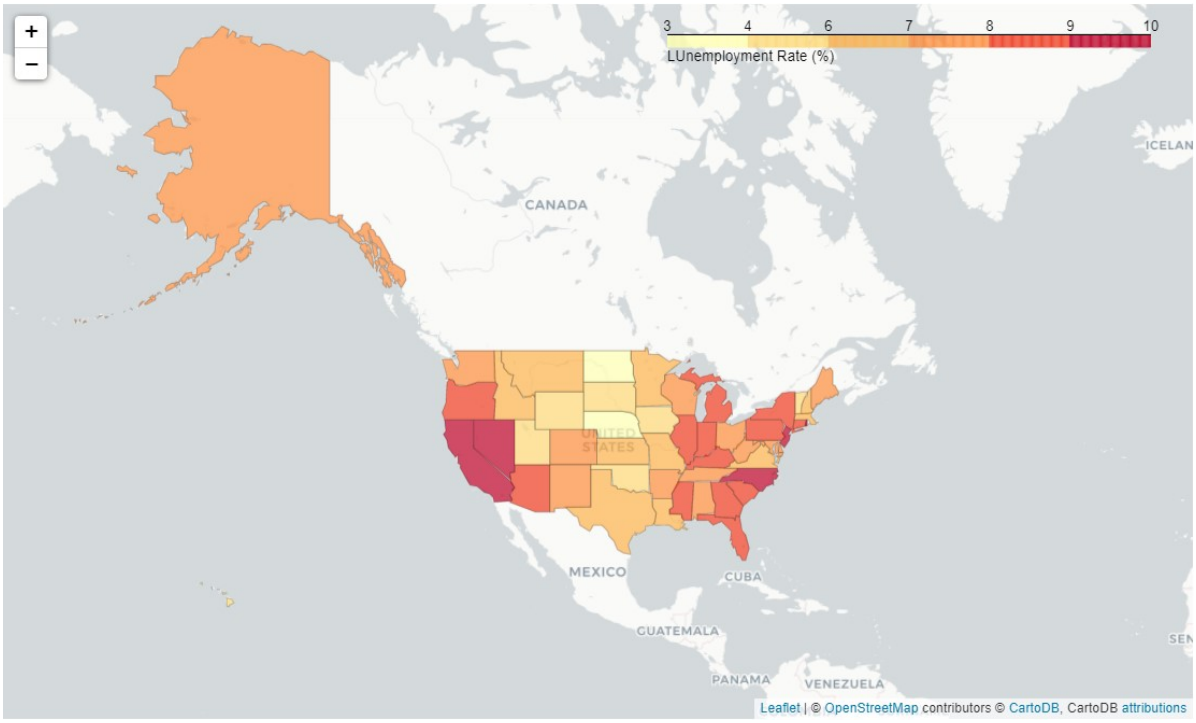


4. Trực quan hóa dữ liệu bản đồ (3 điểm)

- Cho dữ liệu **US_Unemployment_Oct2012.csv** và **us-states.json**, thực hiện các yêu cầu sau :

1. Đọc dữ liệu **US_Unemployment_Oct2012.csv**, hiển thị thông tin chung của dữ liệu bao gồm : head, tail, info, describe (0.75 điểm)
2. Tạo bản đồ thế giới có kiểu **cartodbpositron** với location([48, -102]) và zoom level (zoom_start=3) gợi ý như hình sau : (0.75 điểm)

3. Tạo choropleth map theo 'LUnemployment Rate (%)' của từng bang theo gợi ý như hình sau :(1.5 điểm)



--- Chúc các bạn làm bài tốt ☺ ---