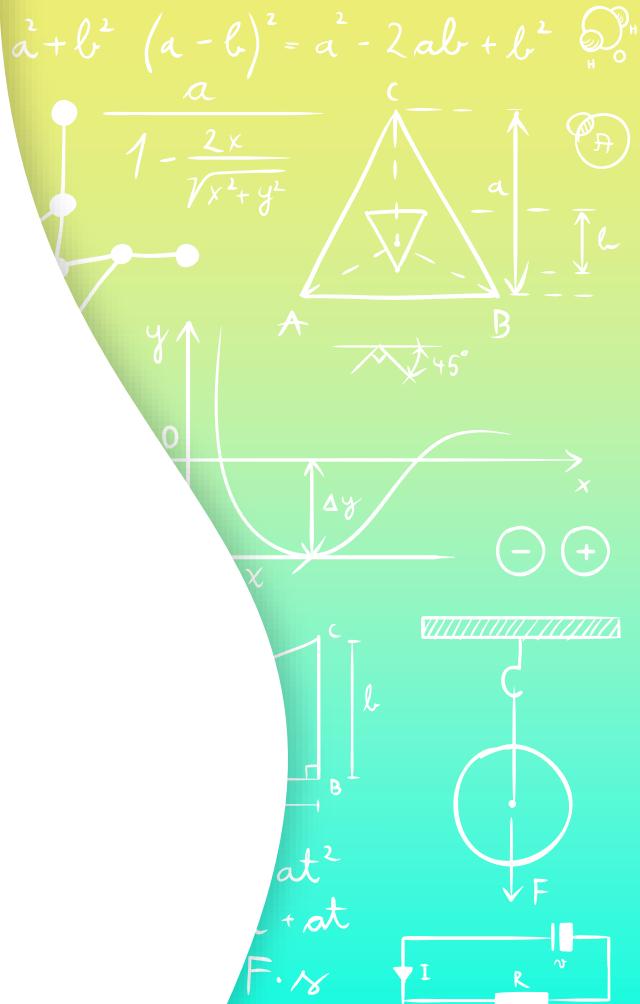


# Programming for Data Science: Data Science Job Salaries Analysis

Analyze how Data Scientists get paid.



# THÀNH VIÊN NHÓM 01

01

Võ Thành Phong

Nhóm trưởng thích giao việc.

03

Bùi Anh Kiệt

AK ông trùm nhắc nhở,  
chúa tể đế xuất.

02

Nguyễn Thị Cẩm Lai

Idol gánh mọi thể loại đồ án.

04

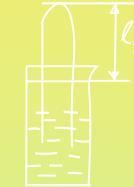
Võ Phi Hùng

Chỉ là một thành viên bình  
thường.

$$T_1 = \ell_1 + 273 = 273 + 60 = 333K, T_2 = \ell_2 + 273 = 298K$$

$$\frac{\ell_1}{\ell_2} = \frac{500}{426} = 1,17$$

$$\Delta \ell_1 = 0,01$$



$$\Delta \ell = \Delta \ell_1 + \Delta \ell_2 = 1 + 0,5 = 1,5 \text{ mm}$$



$$\begin{aligned} &+ 273 \\ &+ 273 \\ &= 327K \end{aligned}$$

$$\Delta T = \Delta T_{\text{at}} \Delta T_0 = 1 + 0,5 = 1,5K$$



# NỘI DUNG TRÌNH BÀY

Những nội dung sẽ có trong bài thuyết trình.

## A. THU THẬP DỮ LIỆU

Trình bày lý do chọn dữ liệu, nguồn và cấu trúc cơ bản của bộ dữ liệu.

## B. KHÁM PHÁ DỮ LIỆU

Thực hiện kiểm tra, xem xét và tiền xử lý dữ liệu.

## C. ĐƯA RA CÂU HỎI VÀ TRẢ LỜI

Thực hiện phân tích, đánh giá và khai thác thông tin từ dữ liệu để trả lời các câu hỏi.

# A. THU THẬP DỮ LIỆU



$$\begin{aligned}
 e &= f^2(x+4gh)^2(s) \cdot (x)^3 \div (gh)^2 - x^2 \\
 f &= gh^2 + (s)(x+2h)^3 \times 4x^2(h)e^3 + x^2 - 2x^2 \\
 g &= x^2 \div (x)(2x)^2 + (hfe)^2 4x^3(3h) \\
 h &= ef^2 - (x)^2 + (3)^2(f)^3 + x(4x)
 \end{aligned}$$

$$a = x(s^1) + (h)(c) + (d)(ef)^2 = x^2$$

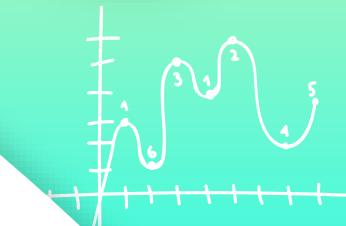
$$(h)(d) \div (s^1)(h^2)(b)^2 = 4x^2 hd$$

$$x^3 \div (x)(x)^2 2x = 2s + 4x$$

$$c^2(h)$$

$$ab = \frac{4x^2 + (ef)^2}{hc \cdot s^2(x)^3}$$

$$x^2 + ab(s)^3 - (x)(s)^1$$



$$\begin{aligned}
 (x)^2 &= ab \\
 (x) &= bc
 \end{aligned}$$

# I. Giới thiệu chủ đề và thông tin tập dữ liệu

## 1. Chủ đề

Tên chủ đề: Data Science Job Salaries

Tạm dịch: Mức lương cho công việc Khoa học dữ liệu

# I. Giới thiệu chủ đề và thông tin tập dữ liệu

## 2. Lý do lựa chọn chủ đề và tập dữ liệu

Hấp dẫn mang tính thực tế, nhận được sự quan tâm lớn từ cộng đồng người theo học và làm việc trong lĩnh vực Khoa học dữ liệu.



# I. Giới thiệu chủ đề và thông tin tập dữ liệu

## 2. Lý do lựa chọn chủ đề và tập dữ liệu

Lợi ích:

- Có cái nhìn tổng quát về sự thay đổi trong lĩnh vực này từ năm 2020 đến hiện tại.
- Nắm bắt được xu thế làm việc và mức lương giữa các ngành nghề đang diễn ra trên thế giới.
- Cung cấp nhiều thông tin bổ ích nhằm đưa ra những định hướng về công việc trong tương lai.

# I. Giới thiệu chủ đề và thông tin tập dữ liệu

## 2. Lý do lựa chọn chủ đề và tập dữ liệu

Nguồn dữ liệu:

- Có sẵn với quy mô đóng góp trên toàn thế giới.
- Được thu thập và cập nhật liên tục từ năm 2020 đến thời điểm hiện tại.
- Được công bố trong phạm vi công cộng, có thể truy cập và tải một cách dễ dàng.



# I. Giới thiệu chủ đề và thông tin tập dữ liệu

## 3. Thông tin tập dữ liệu

- Nguồn gốc: <https://salaries.ai-jobs.net/>
- Giấy phép: Toàn bộ bộ dữ liệu được công bố trong phạm vi công cộng theo giấy phép CC0 (Public Domain).

# I. Giới thiệu chủ đề và thông tin tập dữ liệu

## 3. Thông tin tập dữ liệu

Phương pháp thu thập:

- Được trang web cung cấp biểu mẫu để trả lời khảo sát về công việc và mức lương hiện tại trong lĩnh vực Khoa học dữ liệu. Mọi thông tin được ẩn danh.
- Đóng góp bất cứ lúc nào và chỉ tốn chưa đầy một phút để hoàn thành biểu mẫu.



Work year	Experience level	Employment type
-----	-----	Full-time
The year you want to enter your salary info for.		
Your experience level in the job during the year.		
The type of employment for the role.		
Job title	-----	
The role you worked in during the year. Choose the exact role or the one that comes closest to it.		
Salary (per year)	Salary currency	
-----	US Dollar	-----
Your total compensation/gross salary amount for the year (before deductions like social security, taxes and other contributions).		
The currency of your salary (ideally in USD, EUR or GBP).		
Employee residence	Remote ratio	
-----	-----	-----
Your primary country of residence in during the work year (or your residence for tax purposes if unsure).		
The average amount of remote work (work from home/anywhere/off-site).		
Company location	Company size	
-----	-----	-----
The country of your employer's main office or the branch you were employed/contracted by.		
The average number of people that worked for the company during the year.		
<b>SUBMIT DATA</b>		

Hình biểu mẫu

## II. Tổng quan về cấu trúc tập dữ liệu

### 1. Thời điểm thu thập dữ liệu

Tải ở dạng file csv vào ngày 13/11/2022, gồm 1423 dòng dữ liệu.

### 2. Cấu trúc tập dữ liệu

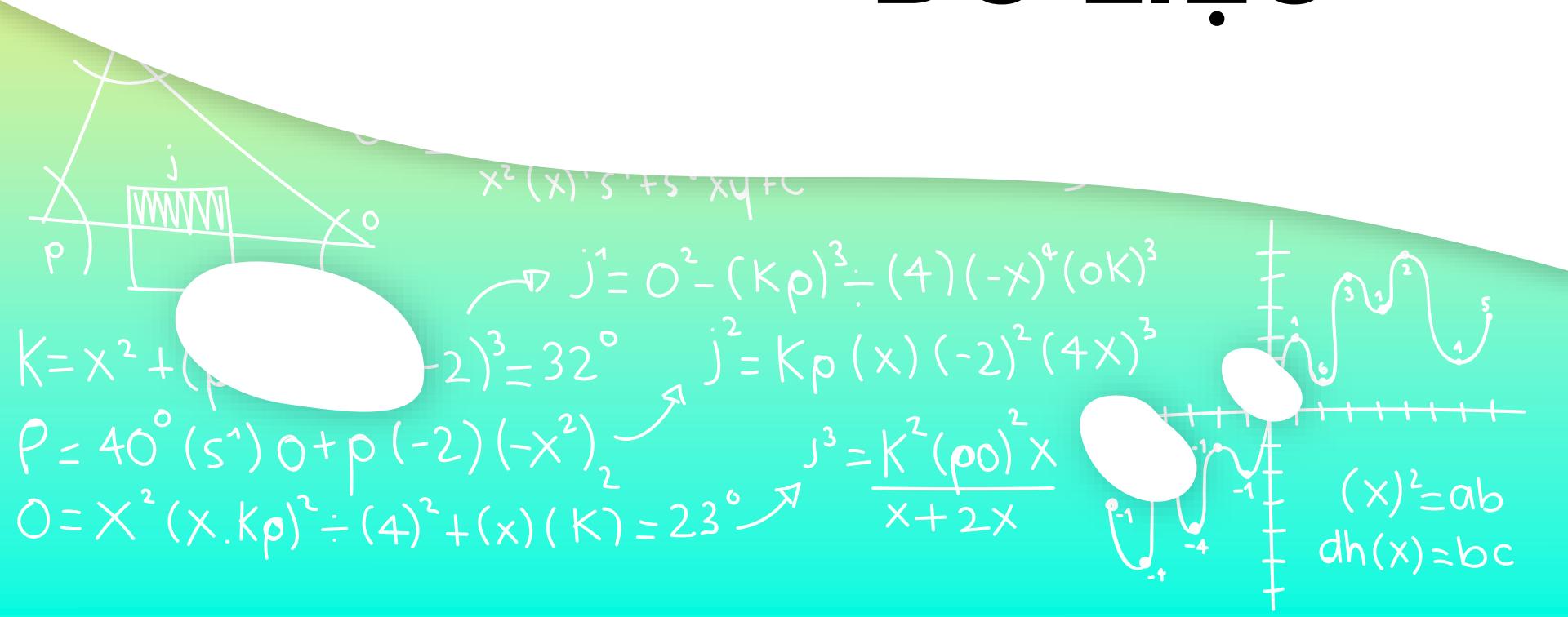
Chứa một bảng duy nhất, gồm 11 thuộc tính.

Tất cả thông tin được cấu trúc và mô tả như sau:

<b>work_year</b>	The year the salary was paid.	
<b>experience_level</b>	The experience level in the job during the year.	EN, MI, SE, EX
<b>employment_type</b>	The type of employment for the role.	PT, FT, CT, FL
<b>job_title</b>	The role worked in during the year.	
<b>salary</b>	The total gross salary amount paid.	
<b>salary_currency</b>	The currency of the salary paid as an ISO 4217 currency code.	
<b>salary_in_usd</b>	The salary in USD (FX rate divided by avg. USD rate for the respective year via fxdata.foorilla.com).	
<b>employee_residence</b>	Employee's primary country of residence in during the work year as an ISO 3166 country code.	
<b>remote_ratio</b>	The overall amount of work done remotely, possible values are as follows.	0, 50, 100
<b>company_location</b>	The country of the employer's main office or contracting branch as an ISO 3166 country code.	
<b>company_size</b>	The average number of people that worked for the company during the year	S, M, L

Cấu trúc tập dữ liệu

# B. KHÁM PHÁ DỮ LIỆU





# Giới thiệu

Thực hiện khám phá dữ liệu đã thu thập bằng cách sử dụng thống kê mô tả để hiểu dữ liệu tốt hơn, tức là để xác định các vấn đề về dữ liệu (dữ liệu bị thiếu giá trị, giá trị không hợp lệ, cột có kiểu dữ liệu không phù hợp để xử lý thêm,...). Thông qua việc khám phá dữ liệu, có thể sẽ phát hiện ra những điểm bất thường, không hợp lý của dữ liệu, từ đó thực hiện tiền xử lý để dữ liệu trở nên rõ ràng và dễ hiểu hơn, phục vụ tốt cho các mục đích khác.

# 1. Đọc dữ liệu và tính số dòng và cột

Đọc file "salaries.csv" vào dataframe `salaries_df` và in ra 5 dòng đầu tiên của dataframe.

```
salaries_df = pd.read_csv('salaries.csv')  
salaries_df.head()
```

	work_year	experience_level	employment_type	job_title	salary	salary_currency	salary_in_usd	employee_residence	remote_ratio	company_location	company_size
0	2022	SE	FT	Data Scientist	205000	USD	205000	US	100	US	M
1	2022	SE	FT	Data Scientist	185000	USD	185000	US	100	US	M
2	2022	SE	FT	Data Scientist	185900	USD	185900	US	0	US	M
3	2022	SE	FT	Data Scientist	129300	USD	129300	US	0	US	M
4	2022	SE	FT	Machine Learning	247500	USD	247500	US	0	US	M

# 1. Đọc dữ liệu và tính số dòng và cột

Tính số dòng và số cột và lưu vào 2 biến num\_rows và num\_cols.

```
num_rows, num_cols = salaries_df.shape  
print(f'Number of rows: {num_rows}\nNumber of columns: {num_cols}')
```

Number of rows: 1423

Number of columns: 11

## **2. Mỗi dòng có ý nghĩa gì? Có vấn đề các dòng có ý nghĩa khác nhau không?**

- Tương ứng với một bản ghi của một người dùng đã điền vào biểu mẫu.
- Chứa các thông tin về thu nhập của người dùng trong một năm cụ thể, vị trí công việc, kinh nghiệm, quốc gia, loại hình công việc, v.v.  
→ Không có dòng nào bị "lạc loài".

### 3. Dữ liệu có các dòng bị lặp không?

Sử dụng các phương thức `duplicated()` và `any()` trên dataframe `salaries_df` và lưu kết quả vào biến `have_duplicated_rows`, có giá trị `True` nếu dữ liệu có các dòng bị lặp và có giá trị `False` nếu ngược lại.

```
have_duplicate_rows = salaries_df.duplicated().any()  
have_duplicate_rows
```

True

→ Có các dòng bị lặp.

Tuy vậy, không bỏ đi các dòng bị lặp vì tính chất của dữ liệu. Hai người khác nhau có thể có các thông tin thu nhập giống nhau trong một năm cụ thể.

# 4. Tỉ lệ giá trị thiếu và thống kê mô tả của từng cột

Đầu tiên, tính tỉ lệ giá trị thiếu của từng cột bằng cách sử dụng phương thức `isnull()` trên dataframe `salaries_df` và tính tổng số giá trị thiếu của từng cột bằng phương thức `sum()`. Cuối cùng, chia số dòng và lưu kết quả vào `missing_ratio`.

```
missing_ratio = salaries_df.isnull().sum()  
missing_ratio = missing_ratio / num_rows  
missing_ratio
```

Tỉ lệ giá trị thiếu đều là 0 nên dữ liệu không có giá trị thiếu.

work_year	0.0
experience_level	0.0
employment_type	0.0
job_title	0.0
salary	0.0
salary_currency	0.0
salary_in_usd	0.0
employee_residence	0.0
remote_ratio	0.0
company_location	0.0
company_size	0.0
dtype:	float64

# 4. Tỉ lệ giá trị thiếu và thống kê mô tả của từng cột

Tính các giá trị thống kê mô tả của các cột numeric bằng phương thức `describe()` trên dataframe `salaries_df`.

```
salaries_df.describe()
```

	work_year	salary	salary_in_usd	remote_ratio
<b>count</b>	1423.000000	1423.000000	1423.000000	1423.000000
<b>mean</b>	2021.735770	236272.560787	124448.492621	62.579058
<b>std</b>	0.547754	1056178.844044	64414.030155	45.776892
<b>min</b>	2020.000000	2324.000000	2324.000000	0.000000
<b>25%</b>	2022.000000	81666.000000	78000.000000	0.000000
<b>50%</b>	2022.000000	130000.000000	123648.000000	100.000000
<b>75%</b>	2022.000000	177550.000000	165400.000000	100.000000
<b>max</b>	2022.000000	30400000.000000	450000.000000	100.000000

# 5. Kiểu dữ liệu của mỗi cột

Sử dụng phương thức dtypes trên dataframe salaries\_df để xem kiểu dữ liệu của mỗi cột. Kết quả được lưu vào series col\_dtypes; series này có index là tên các cột và giá trị là kiểu dữ liệu của các cột tương ứng.

```
col_dtype = salaries_df.dtypes  
col_dtype
```

work_year	int64
experience_level	object
employment_type	object
job_title	object
salary	int64
salary_currency	object
salary_in_usd	int64
employee_residence	object
remote_ratio	int64
company_location	object
company_size	object
dtype: object	

Về mặt kiểu dữ liệu, các thuộc tính của tập dữ liệu này đã ở định dạng phù hợp, nên không cần phải xử lý.

# 6. Xem xét tập giá trị của các thuộc tính phân loại

Xem xét mỗi thuộc tính phân loại có bao nhiêu giá trị phân biệt bằng phương thức `set()`.

```
set(salaries_df['work_year'].to_list())
```

{2020, 2021, 2022}

# 6. Xem xét tập giá trị của các thuộc tính phân loại

Xem xét mỗi thuộc tính phân loại có bao nhiêu giá trị phân biệt bằng phương thức `set()`.

```
set(salaries_df['employment_type'].to_list())
```

```
{'CT', 'FL', 'FT', 'PT'}
```

# 6. Xem xét tập giá trị của các thuộc tính phân loại

Xem xét mỗi thuộc tính phân loại có bao nhiêu giá trị phân biệt bằng phương thức `set()`.

```
set(salaries_df['remote_ratio'].to_list())
```

```
{0, 50, 100}
```

# 6. Xem xét tập giá trị của các thuộc tính phân loại

Xem xét mỗi thuộc tính phân loại có bao nhiêu giá trị phân biệt bằng phương thức `set()`.

```
set(salaries_df['company_size'].to_list())
```

```
{'L', 'M', 'S'}
```

# 6. Xem xét tập giá trị của các thuộc tính phân loại

Xem xét mỗi thuộc tính phân loại có bao nhiêu giá trị phân biệt bằng phương thức `set()`.

```
set(salaries_df['employee_residence'].to_list())
```

```
{'AE', 'AR', 'AT', 'AU', 'AX', 'BE', 'BG', 'BO',
 'BR', 'CA', 'CH', 'CL', 'CN', 'CO', 'CR', 'CZ',
 'DE', 'DK', 'DO', 'DZ', 'EE', 'EG', 'ES', 'FI',
 'FR', 'GB', 'GR', 'HK', 'HN', 'HR', 'HU', 'ID',
 'IE', 'IN', 'IQ', 'IR', 'IT', 'JE', 'JP', 'KE',
 'LU', 'MD', 'MT', 'MX', 'MY', 'NG', 'NL', 'NZ',
 'PH', 'PK', 'PL', 'PR', 'PT', 'RO', 'RS', 'RU',
 'SG', 'SI', 'SK', 'TH', 'TN', 'TR', 'UA', 'US',
 'VN'}
```



# 6. Xem xét tập giá trị của các thuộc tính phân loại

Xem xét mỗi thuộc tính phân loại có bao nhiêu giá trị phân biệt bằng phương thức `set()`.

```
set(salaries_df['company_location'].to_list())  
  
{'AE', 'AL', 'AR', 'AS', 'AT', 'AU', 'BE', 'BO',  
'BR', 'CA', 'CH', 'CL', 'CN', 'CO', 'CZ', 'DE',  
'DK', 'DZ', 'EE', 'EG', 'ES', 'FI', 'FR', 'GB',  
'GR', 'HN', 'HR', 'HU', 'ID', 'IE', 'IL', 'IN',  
'IQ', 'IR', 'IT', 'JP', 'KE', 'LU', 'MD', 'MT',  
'MX', 'MY', 'NG', 'NL', 'NZ', 'PH', 'PK', 'PL',  
'PR', 'PT', 'RO', 'RU', 'SG', 'SI', 'SK', 'TH',  
'TR', 'UA', 'US', 'VN'}
```

# 6. Xem xét tập giá trị của các thuộc tính phân loại

Xem xét mỗi thuộc tính phân loại có bao nhiêu giá trị phân biệt bằng phương thức `set()`.

```
set(salaries_df['job_title'].to_list())
```

{'3D Computer Vision Researcher', 'AI Scientist', 'Analytics Engineer',  
'Applied Data Scientist', 'Applied Machine Learning Scientist', 'Applied  
Scientist', 'BI Analyst', 'BI Data Analyst', 'Big Data Architect', 'Big Data  
Engineer', 'Business Data Analyst', 'Cloud Data Architect', 'Cloud Data  
Engineer', 'Computer Vision Engineer', 'Computer Vision Software Engineer',  
'Data Analyst', 'Data Analytics Consultant', 'Data Analytics Engineer', 'Data  
Analytics Lead', 'Data Analytics Manager', 'Data Architect', 'Data Engineer',  
'Data Engineering Manager', 'Data Manager', 'Data Operations Analyst', 'Data  
Operations Engineer', 'Data Science Consultant', 'Data Science Engineer',  
'Data Science Lead', 'Data Science Manager', 'Data Science Tech Lead', 'Data  
Scientist', 'Data Scientist Lead', 'Data Specialist', 'Director of Data  
Engineering', 'Director of Data Science', 'ETL Developer', 'Finance Data  
Analyst', 'Financial Data Analyst', 'Head of Data', 'Head of Data Science',  
'Head of Machine Learning', 'Lead Data Analyst', 'Lead Data Engineer', 'Lead  
Data Scientist', 'Lead Machine Learning Engineer', 'ML Engineer', 'Machine  
Learning Developer', 'Machine Learning Engineer', 'Machine Learning  
Infrastructure Engineer', 'Machine Learning Manager', 'Machine Learning  
Research Engineer', 'Machine Learning Scientist', 'Marketing Data Analyst',  
'NLP Engineer', 'Power BI Developer', 'Principal Data Analyst', 'Principal  
Data Architect', 'Principal Data Engineer', 'Principal Data Scientist',  
'Product Data Analyst', 'Product Data Scientist', 'Research Engineer',  
'Research Scientist', 'Staff Data Scientist'}

# 6. Xem xét tập giá trị của các thuộc tính phân loại

Nhận xét 1: Cách hiển thị giá trị bằng ký hiệu viết tắt ở một số thuộc tính sau có thể gây khó khăn hoặc nhầm lẫn cho người xem:

experience\_level  
employment\_type  
employee\_residence  
remote\_ratio  
company\_location  
company\_size

# 6. Xem xét tập giá trị của các thuộc tính phân loại

Giải pháp: Thay thế giá trị của các thuộc tính bị viết tắt thành từ đầy đủ nhằm để hiểu rõ ràng về dữ liệu và phục phụ cho quá trình trực quan hóa được rõ ràng hơn.



# 6. Xem xét tập giá trị của các thuộc tính phân loại

Nhận xét 2:

Có rất nhiều thể loại công việc được liệt kê ở đây (65 loại).

Xuất hiện rất nhiều những tiêu đề công việc có thể xếp chung vào một lĩnh vực:

Vd: *Financial Data Analyst, Product Data Analyst, Business Data Analyst,...* có thể gọi chung là *Data Analyst*.

Vd: *Cloud Data Architect, Principal Data Architect, Big Data Architect'*... có thể gọi chung là *Data Architect*.

# 6. Xem xét tập giá trị của các thuộc tính phân loại

Nhận xét 2:

Thậm chí xuất hiện một số công việc được thể hiện ở cả 2 loại tên (điều này có thể xem xét là sự trùng lặp):

Vd: *Machine Learning Engineer* và *ML Engineer* đều được hiểu là *Machine Learning Engineer*.

Vd: *Financial Data Analyst* và *Finance Data Analyst*.

# 6. Xem xét tập giá trị của các thuộc tính phân loại

Giải pháp: Để phù hợp với mục đích khám phá và phân tích dữ liệu, chúng ta sẽ phân chia các tiêu đề công việc vào các thùng tổng quát thích hợp. Ở đây nhóm đã phân loại và chia thành các nhóm chính sau:

- Data Scientist
- Data Engineer
- Data Analyst
- Data Architect
- Machine Learning Engineer
- Machine Learning Scientist
- Computer Vision Engineer
- NLP Engineer
- Research Scientist
- AI Scientist
- Applied Scientist

# 7. Xem xét sự phân bố giá trị của các cột dữ liệu dạng số

salaries\_df.dtypes

work_year	int64
experience_level	object
employment_type	object
job_title	object
salary	int64
salary_currency	object
salary_in_usd	int64
employee_residence	object
remote_ratio	object
company_location	object
company_size	object
dtype:	object

Sau các bước tiền xử lý ở trên có 3 cột dữ liệu kiểu số là:

work\_year  
salary  
salary\_in\_usd

# 7. Xem xét sự phân bố giá trị của các cột dữ liệu dạng số

Thực hiện thống kê trên 3 cột này và lưu vào một dataframe với các dòng là đại diện cho các giá trị như sau:

- Tỉ lệ % (từ 0 đến 100) các giá trị thiếu (*missing\_ratio*).
- Giá trị *min* (*min*).
- Giá trị *lower quartile* (phân vị 25) (*lower\_quartile*).
- Giá trị *median* (phân vị 50) (*median*).
- Giá trị *upper quartile* (phân vị 75) (*upper\_quartile*).
- Giá trị *max* (*max*).

```
numeric_df=salaries_df.copy()
numeric_df=numeric_df[['work_year', 'salary', 'salary_in_usd']]
columns=list(numeric_df.columns)
titles=['missing_ratio', 'min', 'lower_quartile', 'median', 'upper_quartile', 'max']
arrays=numeric_df.to_numpy()
num_col_dict={}
num_col_vals=[]
for i in range(len(columns)):
    temp=[]
    temp.append(((sum(np.isnan(arrays[:,i]))/arrays.shape[0])*100).round(3))
    temp.append(np.nanmin(arrays[:,i],axis=0).round(1))
    temp.append(np.nanpercentile(arrays[:,i],25,axis=0).round(3))
    temp.append(np.nanpercentile(arrays[:,i],50,axis=0).round(3))
    temp.append(np.nanpercentile(arrays[:,i],75,axis=0).round(3))
    temp.append(np.nanmax(arrays[:,i],axis=0).round(3))
    num_col_vals.append(temp)
for i in range(len(columns)):
    num_col_dict[columns[i]]=num_col_vals[i]
    num_col_dict['titles']=titles
numeric_info_df=pd.DataFrame(num_col_dict).set_index('titles')
numeric_info_df
```

# 7. Xem xét sự phân bố giá trị của các cột dữ liệu dạng số

	work_year	salary	salary_in_usd
<b>titles</b>			
<b>missing_ratio</b>	0.0	0.0	0.0
<b>min</b>	2020.0	2324.0	2324.0
<b>lower_quartile</b>	2022.0	81666.0	78000.0
<b>median</b>	2022.0	130000.0	123648.0
<b>upper_quartile</b>	2022.0	177550.0	165400.0
<b>max</b>	2022.0	30400000.0	450000.0

# **7. Xem xét sự phân bố giá trị của các cột dữ liệu dạng số**

Nhận xét:

Các cột dạng số không có giá trị thiếu, có thể là do cách thu thập thông tin thật sự hiệu quả từ trang web khi thu thập bằng cách cho những người tham gia khảo sát trên toàn thế giới điền biểu mẫu với số lượng thuộc tính vừa phải, và điền một cách ẩn danh do đó những thông tin cơ bản nhưng quan trọng như năm bắt đầu làm việc hay tiền lương thì người dùng có thể săn sàng điền một cách nhanh chóng.

Cách biệt giữa giá trị min và max của cột salary khá lớn là do sự khác biệt về đơn vị tiền tệ ở mỗi quốc gia. Do đó việc có cột salary\_in\_usd là cần thiết cho quá trình trực quan hóa.

$$\begin{aligned}ut &+ \frac{1}{2}at^2 \\r &= u + at \\w &= F \cdot r\end{aligned}$$

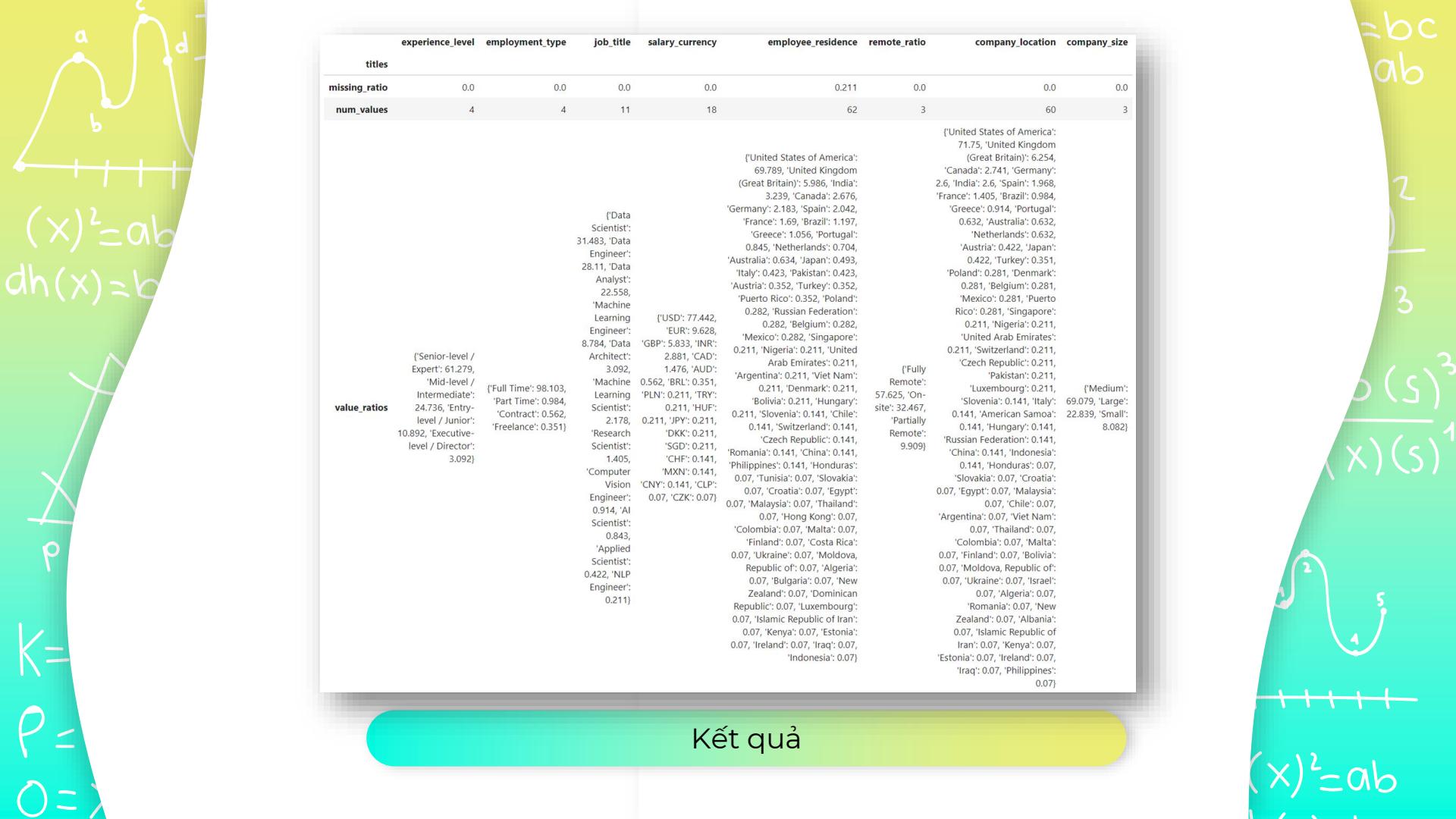
# 8. Xem xét sự phân bố giá trị của các cột dữ liệu không phải dạng số

Thực hiện thống kê và lưu vào một dataframe với các dòng là đại diện cho các giá trị như sau:

- Tỉ lệ % (từ 0 đến 100) các giá trị thiếu (*missing\_ratio*).
- Số lượng các giá trị khác nhau (không xét giá trị thiếu) (*num\_values*).
- Tỉ lệ % (từ 0 đến 100) của mỗi giá trị được sort theo tỉ lệ % giảm dần (không xét giá trị thiếu, tỉ lệ là tỉ lệ so với số lượng các giá trị không thiếu): dùng dictionary để lưu, key là giá trị, value là tỉ lệ % (*value\_ratios*).

$$\begin{aligned}ut + \frac{1}{2}at^2 \\v = u + at \\w = F \cdot t\end{aligned}$$

```
categorical_df=salaries_df.select_dtypes(exclude=['int64'])
col_names=list(categorical_df.columns)
titles=["missing_ratio", "num_values",
"value_ratios"]
values_list=[]
dic={}
for i in col_names:
    df=categorical_df.copy()=[[i]]
    size=len(df)
    new=list(df[i].values)
    df[i]=new
    temp=[]
    temp.append(((df[i].isnull().sum()/size
)*100).round(3))
    df.dropna(inplace=True)
    vals=list(df[i].values)
    num_vals=list(set(vals))
    temp.append(len(num_vals))
    temp2={}
    new=[(df[i].value_counts())[j]/len(vals)
        *100).round(3) for j in num_vals]
    temp2={num_vals[j]:new[j] for j in
range(len(num_vals))}
    temp2_sort={k:v for k,v in
sorted(temp2.items(), key= lambda
item:item[1], reverse=True)}
    temp.append(temp2_sort)
    values_list.append(temp)
dic={col_names[i]:values_list[i] for i in
range(len(col_names))}
dic['titles']=titles
categorical_info_df=pd.DataFrame(dic).set_i
ndex('titles')
categorical_info_df
```



# 8. Xem xét sự phân bố giá trị của các cột dữ liệu không phải dạng số

Nhận xét:

- Hầu như cũng gần như không có giá trị thiếu ở các cột không phải dạng số này.
- Ta thấy phần trăm của loại làm việc từ xa toàn thời gian (Fully Remote) là cao nhất có thể là do bộ dữ liệu nhóm đang dùng là trong khoảng thời gian từ năm 2020 đến năm 2022 do đó ảnh hưởng của dịch covid-19 đã làm cho phần trăm số lượng người làm việc từ xa tăng lên đáng kể.

$$\begin{aligned}ut + \frac{1}{2}at^2 \\v = u + at \\w = F \cdot \Delta t\end{aligned}$$

$$e = f^2(x+4gh)^2(s) \cdot (\wedge)^3 \div (gh)^2 - x^2$$

$$f = gh^2 + (s)(x+2h)^3 \times 4x^2(hc)^3 + x^2 - 2x^2$$

$$g = x^2 \div (x)(2x)^2 + (hf)^2 4x^3(3h)(f)^2(e)^2 + x^2 4s^2$$

$$h = ef g^2 - (x)^2 + (3)^2(f)^3 + x(4x)^2$$

$$(d)(ef)^2 = x^2$$

$$(b)^2 = \frac{4x^2 hd}{2s+4x}$$

$$ab = \frac{4x^2 + (ef)^2}{hc \cdot s^2(x)_3}$$

# ĐƯA RA CÂU HỎI VÀ TRẢ LỜI

# Giới thiệu

- Suy nghĩ câu hỏi có thể hỏi từ bộ dữ liệu và tìm lời giải đáp một cách trực quan sinh động nhất thông qua các bước tiền xử lý và phân tích dữ liệu.
- Lưu ý nhỏ: Các insight trong phần này hoàn toàn dựa trên việc phân tích người đã khảo sát, nên có thể sẽ không đại diện được cho cả cộng đồng người trong lĩnh vực Khoa Học Dữ Liệu.

# Câu hỏi 1

**Mức lương trung bình của công việc khoa học dữ liệu biến động như thế nào trong những năm gần đây ở trên thế giới?**

Ý nghĩa khi trả lời được câu hỏi: Có cái nhìn tổng quát về sự thay đổi của mức lương trung bình trong lĩnh vực Khoa học dữ liệu trong ba năm (2020, 2021, 2022). Từ đó, thấy được xu hướng phát triển và lý giải được nguyên nhân thực tế làm tăng/giảm của mức lương.

Tiền xử lý: Không cần thiết cho câu hỏi này.

# Câu hỏi 1

**Mức lương trung bình của công việc khoa học dữ liệu biến động như thế nào trong những năm gần đây ở trên thế giới?**

Phân tích dữ liệu để trả lời cho các câu hỏi:

Bước 1: Tạo dataframe 'salary\_timeline' gồm 2 thuộc tính: 'work\_year' (để lưu 3 năm: 2020, 2021, 2022), 'mean\_salary\_in\_usd' (lưu mức lương trung bình của năm tương ứng).

# Câu hỏi 1

**Mức lương trung bình của công việc khoa học dữ liệu biến động như thế nào trong những năm gần đây ở trên thế giới?**

Phân tích dữ liệu để trả lời cho các câu hỏi:

Bước 2:

- Tiến hành trực quan hóa bằng *lineplot* (thư viện seaborn) trong đó: trực hoành là năm (work\_year) và trực tung sẽ là mức lương trung bình (salary\_in\_usd)
- Điều chỉnh các tham số và cài đặt các label/title..., sao cho hình ảnh trực quan được rõ ràng đẹp mắt.

# Câu hỏi 1

**Mức lương trung bình của công việc khoa học dữ liệu biến động như thế nào trong những năm gần đây ở trên thế giới?**

Code bước 1:

```
salary_timeline =  
pd.DataFrame(salaries_df.groupby("work_year")["salary_in_usd"]  
.mean())  
salary_timeline = salary_timeline.reset_index()  
salary_timeline.rename(columns =  
{'salary_in_usd':'mean_salary_in_usd'}, inplace = True)  
salary_timeline["work_year"].replace({2020:"2020",2021:"2021",  
2022:"2022"},inplace=True)
```

# Câu hỏi 1

**Mức lương trung bình của công việc khoa học dữ liệu biến động như thế nào trong những năm gần đây ở trên thế giới?**

Hãy nhìn vào dataframe chứa dữ liệu cần thiết cho câu hỏi này:

work_year	mean_salary_in_usd
-----------	--------------------

0	2020	92644.413333
1	2021	93616.787611
2	2022	132784.741533

# Câu hỏi 1

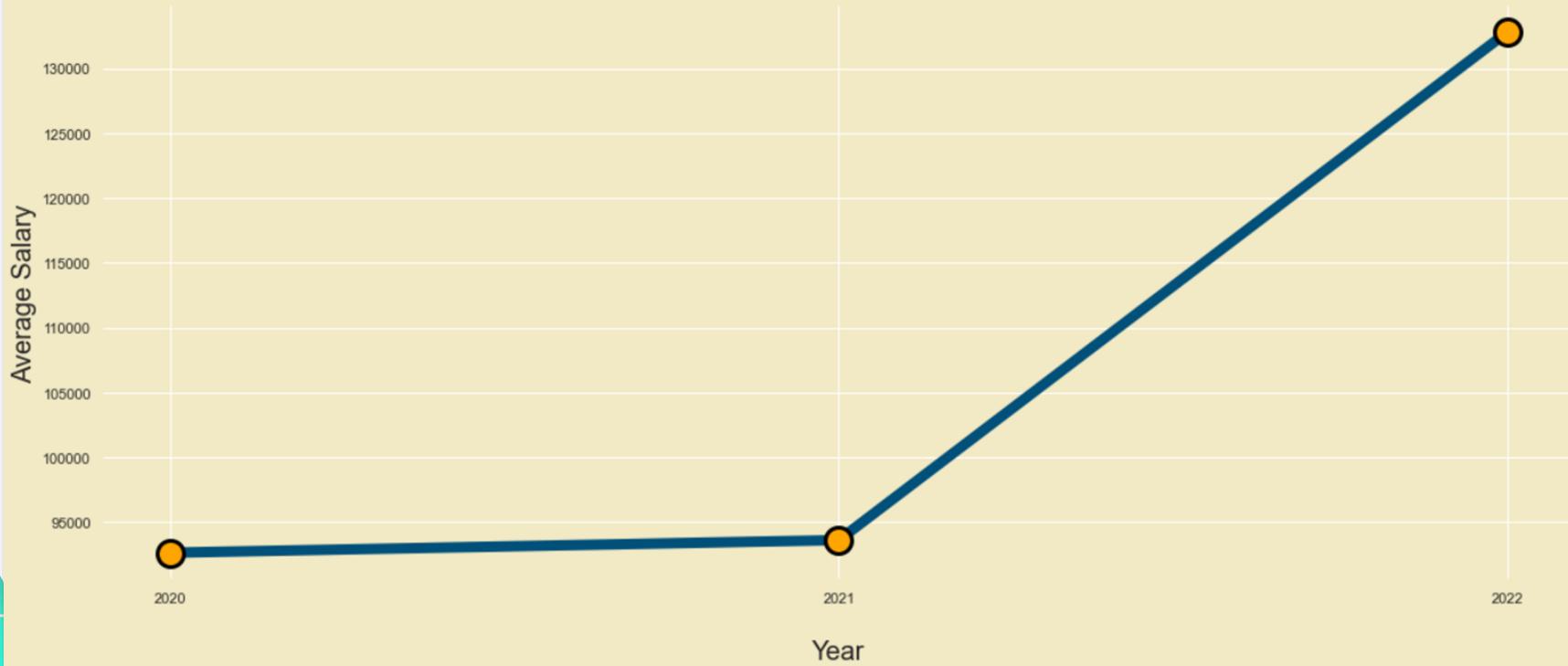
Code bước 2:

```
sns.set(rc={"axes.facecolor":"#F2EAC5","figure.facecolor":"#F2EAC5"})
palette =
["#11264e", "#00507A", "#026e90", "#008b99", "#6faea4", "#fcdbcb", "#FEE08B", "#faa96e", "#f36b3b", "#ef3f28", "#CC0028"]
plt.subplots(figsize=(20,8))
p=sns.lineplot(x=salary_timeline["work_year"]
,y=salary_timeline["mean_salary_in_usd"],data=salary_timeline,color=pale
tte[1],marker="o", linewidth=8, markersize=20, markerfacecolor="orange", mar
keredgecolor="black", markeredgewidth=3)
p.axes.set_title("\nSự thay đổi mức lương trung bình qua các năm\n",
fontsize=25)
p.axes.set_xlabel("\nYear", fontsize=20)
p.axes.set_ylabel("Average Salary", fontsize=20)
sns.despine(left=True, bottom=True)
plt.show()
```

$$ab + b^2$$



## Sự thay đổi mức lương trung bình qua các năm



$$y = -2x$$



# Câu hỏi 1

**Mức lương trung bình của công việc Khoa học dữ liệu biến động như thế nào trong những năm gần đây ở trên thế giới?**

Nhận xét:

- Mức lương trung bình trong lĩnh vực Khoa học dữ liệu có xu hướng tăng từ năm 2020 đến năm 2022:
- Tăng nhẹ trong khoảng từ 2020 đến 2021: từ 92644 -> 93616 (usd)
- Tăng mạnh trong khoảng từ 2021 đến 2022: từ 93616 -> 132784 (usd)
- Điều này cho thấy lĩnh vực Khoa học dữ liệu đang ngày càng đóng vai trò quan trọng, trở thành xu thế nghề nghiệp được ưa chuộng, nguồn nhân lực được các công ty trên thế giới săn sàng săn đón với mức lương hậu hĩnh.

# Câu hỏi 2

**Đâu là mảnh đất lý tưởng cho việc lựa chọn nơi làm việc cho các ngành thuộc lĩnh vực Khoa học dữ liệu?**

Ý nghĩa khi trả lời được câu hỏi: Giúp cho người tìm việc chọn được các quốc gia nơi mà lĩnh vực Khoa học dữ liệu đang phát triển, sẵn sàng chi trả cho nguồn nhân lực với mức lương hấp dẫn.

# Câu hỏi 2

**Đâu là mảnh đất lý tưởng cho việc lựa chọn nơi làm việc cho các ngành thuộc lĩnh vực Khoa học dữ liệu?**

Tiền xử lý:

- Xử lý dữ liệu để có thể trực quan hóa kết quả trên bản đồ thế giới (sử dụng thư viện px.choropleth)
- Để sử dụng được thư viện px.choropleth, để dữ liệu có thể map lên được bản đồ thế giới, thì tên của quốc gia phải ở dạng ISO (Tổ chức tiêu chuẩn hóa quốc tế), vì vậy cần phải chuyển đổi giá trị cột 'company\_location' về dạng mã ISO.

# Câu hỏi 2

**Đâu là mảnh đất lý tưởng cho việc lựa chọn nơi làm việc cho các ngành thuộc lĩnh vực Khoa học dữ liệu?**

Code tiền xử lý:

```
country_names =  
coco.convert(names=salaries_df['company_location'], to="ISO3")  
salaries_df['company_location'] = country_names
```

# Câu hỏi 2

**Đâu là mảnh đất lý tưởng cho việc lựa chọn nơi làm việc cho các ngành thuộc lĩnh vực Khoa học dữ liệu?**

Code tiền xử Sau khi chuyển tên quốc gia về mã ISO ta có:

```
array(  
    ['USA', 'HUN', 'GBR', 'CHE', 'DEU', 'AUT', 'SVK', 'IND', 'ESP',  
    'FRA', 'CAN', 'MEX', 'BRA', 'AUS', 'SGP', 'PRT', 'NGA', 'CZE',  
    'TUR', 'PRI', 'FIN', 'ASM', 'THA', 'NLD', 'GRC', 'DNK', 'BOL',  
    'PHL', 'ALB', 'ARG', 'BEL', 'IDN', 'EGY', 'ITA', 'ARE', 'IRL',  
    'LUX', 'SVN', 'MYS', 'EST', 'POL', 'HND', 'PAK', 'JPN', 'DZA',  
    'ROU', 'IRQ', 'RUS', 'UKR', 'CHN', 'KEN', 'COL', 'NZL', 'IRN',  
    'CHL', 'MDA', 'VNM', 'HRV', 'ISR', 'MLT'], dtype=object)
```

# Câu hỏi 2

**Đâu là mảnh đất lý tưởng cho việc lựa chọn nơi làm việc cho các ngành thuộc lĩnh vực Khoa học dữ liệu?**

Phân tích dữ liệu để trả lời cho các câu hỏi:

Bước 1: Tạo dataframe 'average\_salary' gồm có 2 thuộc tính là 'company\_location' và 'mean\_salary\_in\_usd':

- 'company\_location': thể hiện mã các quốc gia theo ISO.
- 'mean\_salary\_in\_usd': thể hiện giá trị trung bình về lương (trong vòng 3 năm) của các quốc gia tương ứng.

# Câu hỏi 2

**Đâu là mảnh đất lý tưởng cho việc lựa chọn nơi làm việc cho các ngành thuộc lĩnh vực Khoa học dữ liệu?**

Phân tích dữ liệu để trả lời cho các câu hỏi:

Bước 2:

- Tiến hành trực quan hóa bằng biểu đồ choropleth (thư viện plotly.express) trong đó: giá trị mức lương trung bình (mean\_salary\_in\_usd) của từng quốc gia (company\_location) sẽ được hiển thị theo màu sắc trên bảng đồ thế giới.
- Điều chỉnh các tham số và thiết kế các layout sao cho hình ảnh trực quan được rõ ràng và đẹp mắt.

# Câu hỏi 2

**Đâu là mảnh đất lý tưởng cho việc lựa chọn nơi làm việc cho các ngành thuộc lĩnh vực Khoa học dữ liệu?**

Code bước 1:

```
salary_location_df = salaries_df.groupby(['salary_in_usd',  
'company_location']).size().reset_index()  
average_salary =  
salary_location_df.groupby('company_location').mean().reset_index()  
average_salary.rename(columns = {'salary_in_usd':'mean_salary_in_usd'},  
inplace = True)  
average_salary =  
average_salary[['company_location','mean_salary_in_usd']]
```

	company_location	mean_salary_in_usd						
0	ALB	2324.000000	20	EST	31548.000000	40	MLT	28369.000000
1	ARE	100000.000000	21	FIN	63096.000000	41	MYS	40000.000000
2	ARG	50000.000000	22	FRA	60438.250000	42	NGA	86666.666667
3	ASM	34026.500000	23	GBR	87783.480000	43	NLD	70507.111111
4	AUS	83578.250000	24	GRC	50244.000000	44	NZL	125000.000000
5	AUT	71377.666667	25	HND	20000.000000	45	PAK	13333.333333
6	BEL	76895.250000	26	HRV	45618.000000	46	PHL	50000.000000
7	BOL	7500.000000	27	HUN	26778.500000	47	POL	65595.000000
8	BRA	34396.454545	28	IDN	34410.500000	48	PRI	167500.000000
9	CAN	110812.200000	29	IND	27130.031250	49	PRT	47285.222222
10	CHE	76848.666667	30	IRL	68355.000000	50	ROU	60000.000000
11	CHL	40038.000000	31	IRN	4000.000000	51	RUS	157500.000000
12	CHN	71665.500000	32	IRQ	100000.000000	52	SGP	54143.000000
13	COL	21844.000000	33	ISR	119059.000000	53	SVK	12619.000000
14	CZE	35207.333333	34	ITA	36366.500000	54	SVN	63831.000000
15	DEU	76706.062500	35	JPN	114127.333333	55	THA	15000.000000
16	DNK	45558.000000	36	KEN	9272.000000	56	TUR	19058.000000
17	DZA	100000.000000	37	LUX	43942.666667	57	UKR	13400.000000
18	EGY	22800.000000	38	MDA	18000.000000	58	USA	151170.680912
19	ESP	50149.000000	39	MEX	31592.500000	59	VNM	4000.000000

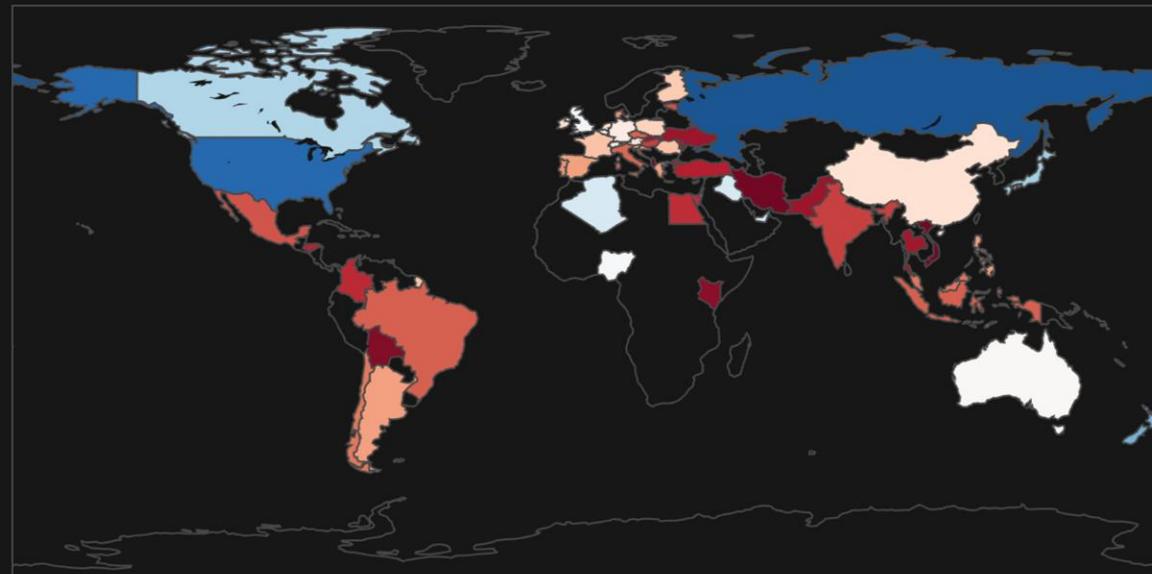
# Câu hỏi 2

**Đâu là mảnh đất lý tưởng cho việc lựa chọn nơi làm việc cho các ngành thuộc lĩnh vực Khoa học dữ liệu?**

Code bước 2:

```
fig = px.choropleth(locations=average_salary['company_location'],
                     color=average_salary['mean_salary_in_usd'],
                     color_continuous_scale=px.colors.sequential.RdBu,
                     template='plotly_dark')
fig.update_layout(font = dict(size=17,family="Courier new"))
fig.update_layout(
    title="Average Salary by Company Location", title_x=0.5,
    font=dict(family="Rubik", size=18))
fig.show()
```

## Average Salary by Company Location



color

150k

100k

50k



# Câu hỏi 2

**Đâu là mảnh đất lý tưởng cho việc lựa chọn nơi làm việc cho các ngành thuộc lĩnh vực Khoa học dữ liệu?**

Nhận xét:

- Dựa vào hình ảnh được trực quan ta có thể thấy, các quốc gia có mức lương trung bình rất cao đó là: Nga (157.5k), Mỹ (151.1707k), New Zealand (125k), Nhật Bản (114.1273k), Canada (110.8122k).

# Câu hỏi 2

**Đâu là mảnh đất lý tưởng cho việc lựa chọn nơi làm việc cho các ngành thuộc lĩnh vực Khoa học dữ liệu?**

Nhận xét:

- Đây đều là những quốc gia phát triển hàng đầu thế giới về mọi lĩnh vực và các quốc gia này sẵn sàng chi trả mức lương cho nguồn nhân lực trong lĩnh vực Khoa học dữ liệu rất hậu hĩnh. Điều đó chứng tỏ Khoa học dữ liệu là một lĩnh vực nghề nghiệp quan trọng trong sự phát triển của đất nước và người lao động có thể cân nhắc lựa chọn các quốc gia này để phát triển sự nghiệp và có mức thu nhập tốt.

# Câu hỏi 3

**Các công việc nào được làm nhiều nhất ở mỗi quốc gia?**

Ý nghĩa khi trả lời được câu hỏi:

- Người đọc ở một quốc gia cụ thể có thể định hướng về nghề nghiệp tương lai cho bản thân.
- Có thông tin về cơ hội việc làm ở trong nước hay các quốc gia khác.
- So sánh với các năm 2020 và 2021 để xem xét sự thay đổi số lượng nhân viên trong một công việc cụ thể để biết liệu công việc có bị bão hòa hay vẫn còn cần nhu cầu nhân viên.

Tiền xử lý: Không cần thiết cho câu hỏi này.

# Câu hỏi 3

**Các công việc nào được làm nhiều nhất ở mỗi quốc gia?**

Phân tích dữ liệu để trả lời cho các câu hỏi:

- Tạo một DataFrame với các dòng là chỉ số năm làm việc, các cột là các quốc gia, các giá trị là những dictionary với keys là tên ngôn ngữ và values là tỉ lệ phần trăm người sử dụng ngôn ngữ đó(tính trong một quốc gia) được sắp xếp giảm dần.
- Do có nhiều quốc gia nên trực quan hóa dạng biểu đồ sẽ quá phức tạp và không rõ ràng. Nên đối với câu hỏi này việc trực quan được thể hiện ngay trên dataframe để rõ ràng và dễ nhìn hơn.

```

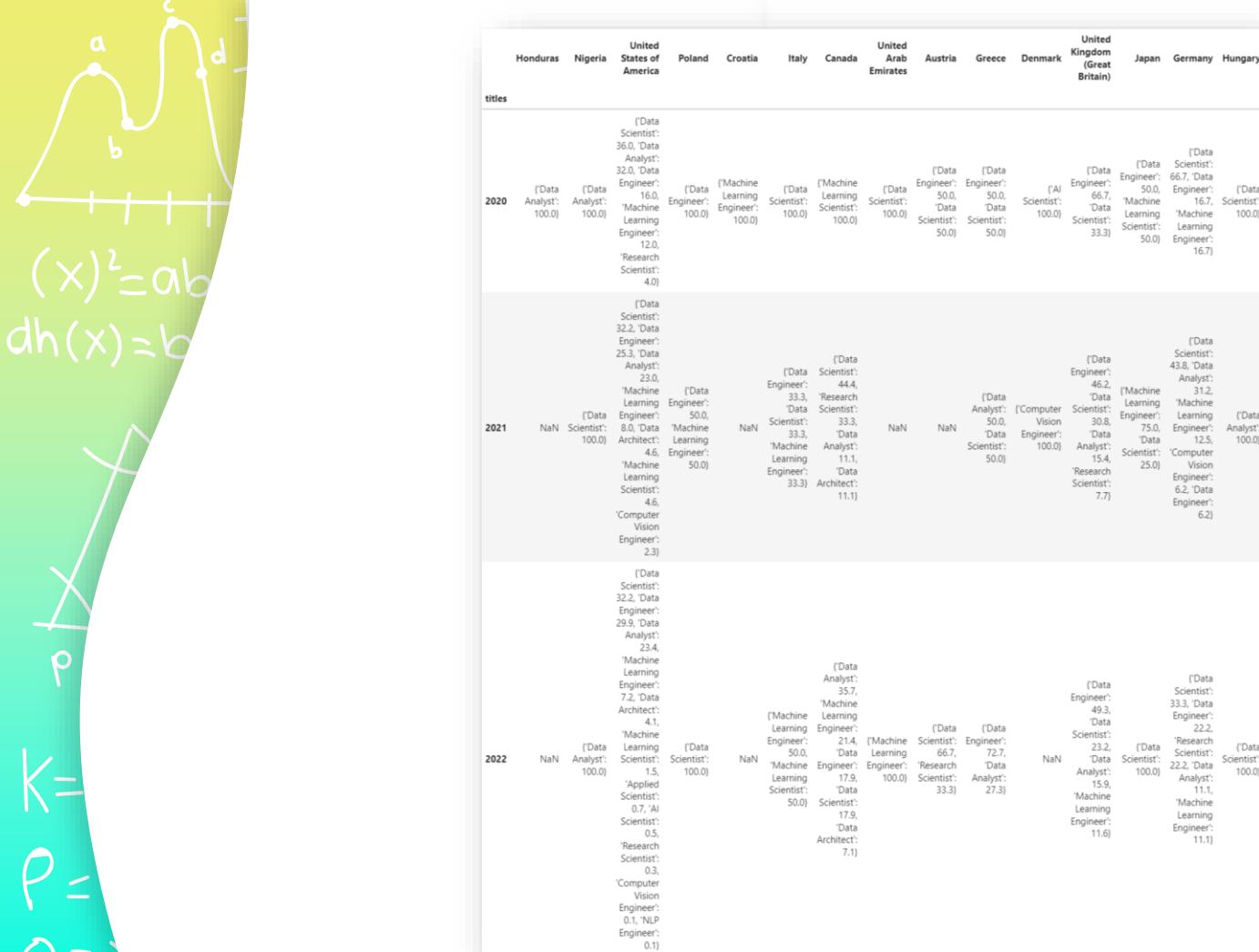
pd.set_option('display.max_colwidth', -1)
pd.set_option('display.max_columns', None)
pd.set_option('display.max_rows', None)
def get_country_jobs(data, year):
    country_col, job_col='employee_residence', 'job_title'
    df=data.copy()[[country_col,job_col]]
    #loại các dòng có quốc gia hoặc công việc là nan
    df.dropna(inplace=True)
    country_names=list(set(list(df[country_col].values)))
    country_vals=[]
    for c in country_names:
        sub_df=df[df[country_col]==c]
        jobs=list(sub_df[job_col].values)
        jobs=[i.split(';') for i in jobs]
        jbs=np.array([j for i in jobs for j in i])
        n,c=np.unique(jbs,return_counts=True)[0],
        np.unique(jbs,return_counts=True)[1]
        c=((c/jbs.shape[0])*100).round(1)
        n,c=list(n),list(c)
        dic_jbs=dict(zip(n,c))
        dic_jbs_sort={k:v for k,v in

```

```

sorted(dic_jbs.items(), key= lambda item:item[1],
reverse=True)}
    temp=[]
    temp.append(dic_jbs_sort)
    country_vals.append(temp)
    country_jobs_dict={country_names[i]:country_vals[i] for i in range(len(country_names))}}
    country_jobs_dict['titles']=year
    country_jobs_df=pd.DataFrame(country_jobs_dict).set_index('titles')
    return country_jobs_df
df2020=get_country_jobs(salaries_df[salaries_df['work_year']==2020], '2020')
df2021=get_country_jobs(salaries_df[salaries_df['work_year']==2021], '2021')
df2022=get_country_jobs(salaries_df[salaries_df['work_year']==2022], '2022')
country_jobs_df=pd.concat([df2020, df2021, df2022], axis=0)

```



# Câu hỏi 3

**Các công việc nào được làm nhiều nhất ở mỗi quốc gia?**

Nhận xét:

Người đọc có thể xem kết quả ở cột quốc gia mà người đọc quan tâm để có thể đưa ra nhận định cho bản thân về công việc mong muốn ở quốc gia mong muốn.

# Câu hỏi 3

**Các công việc nào được làm nhiều nhất ở mỗi quốc gia?**

Nhận xét:

Tuy nhiên hạn chế của bộ dữ liệu ảnh hưởng đến tính tổng quát của câu hỏi như sau:

- Hạn chế về việc số người ở một quốc gia nào đó tham gia khảo sát mà sẽ làm cho một số quốc gia không đủ số liệu của cả 3 năm 2020, 2021, 2022.
- Điều này phần nào làm mất tính tổng quát khi trả lời câu hỏi được đề ra nhưng người đọc vẫn có thể xem qua những số liệu hiện có của quốc gia đó để tham khảo (do các năm khá gần nhau, nên số liệu vẫn có thể có giá trị).

# Câu hỏi 3

**Các công việc nào được làm nhiều nhất ở mỗi quốc gia?**

Nhận xét:

Nếu trang web thu thập dữ liệu có thể đảm bảo các quốc gia có người khảo sát năm trước sẽ quay lại khảo sát năm sau thì số liệu thống kê cho câu hỏi sẽ được thống kê một cách đầy đủ.

# Câu hỏi 3

**Các công việc nào được làm nhiều nhất ở mỗi quốc gia?**

Nhận xét:

Vậy theo nhóm tại sao câu hỏi vẫn rất có giá trị?

- Chỉ có một số ít quốc gia không đủ số liệu thống kê cho 3 năm, nhưng có rất nhiều quốc gia lớn mà ở đó cơ hội việc làm cao có đầy đủ thống kê cho 3 năm.
- Việc thiếu dữ liệu này hoàn toàn có thể được trang web khắc phục bằng cách mở rộng quy mô khảo sát đến các quốc gia thông qua đường Internet.
- Như đã đề cập ở trên nếu quốc gia nào chỉ có thống kê cho một hoặc hai năm thì số liệu vẫn có giá trị khảo sát do các năm khá gần nhau.

# Câu hỏi 3

**Các công việc nào được làm nhiều nhất ở mỗi quốc gia?**

Nhận xét:

```
country_jobs_df['Viet Nam']
```

```
titles
2020    NaN
2021    {'Data Analyst': 33.3, 'Data Scientist': 33.3, 'Machine Learning Scientist': 33.3}
2022    NaN
Name: Viet Nam, dtype: object
```

Theo khảo sát từ trang AI Jobs thì ở Việt Nam năm 2021 có 3 công việc chính cho ngành khoa học dữ liệu là: Data Analyst, Data Scientist, Machine Learning Scientist với tỉ lệ ngang bằng nhau.



$$T_1 = t_1 + 273 = 273 + 60 = 333K, T_2 = t_2 + 273 = 298K$$

# Câu hỏi 4

**Tỉ lệ làm việc cho công ty nước ngoài của mỗi công việc (một nhân viên được xem là làm việc cho công ty nước ngoài nếu có employee\_residence khác company\_location)?**

Ý nghĩa khi trả lời được câu hỏi:

- Ý định làm việc cho công ty nước ngoài của người đọc cho một công việc cụ thể có khả thi hay không?
- Những công việc nào có môi trường làm việc ở các nước rộng lớn bằng cách xem công việc nào có tỉ lệ làm việc cho công ty nước ngoài cao nhất.

Tiền xử lý: Không cần thiết cho câu hỏi này.



$$T_1 = \ell_1 + 273 = 273 + 60 = 333K, T_2 = \ell_2 + 273 = 298K$$

# Câu hỏi 4

**Tỉ lệ làm việc cho công ty nước ngoài của mỗi công việc (một nhân viên được xem là làm việc cho công ty nước ngoài nếu có employee\_residence khác company\_location)?**

Phân tích dữ liệu để trả lời cho các câu hỏi:

Bước 1:

- Tạo một pandas.Series với index là tên các công việc (job\_title), values là phần trăm số lượng các người tham gia khảo sát làm cho công ty nước ngoài.
- Sắp xếp theo values giảm dần.



$$l + 0,5 = 1,5 \text{ mm}$$

$$T_1 = t_1 + 273 = 273 + 60 = 333K, T_2 = t_2 + 273 = 298K$$

→ T

# Câu hỏi 4

**Tỉ lệ làm việc cho công ty nước ngoài của mỗi công việc (một nhân viên được xem là làm việc cho công ty nước ngoài nếu có employee\_residence khác company\_location)?**

Phân tích dữ liệu để trả lời cho các câu hỏi:

Bước 2:

- Vẽ barchart để minh họa phần trăm làm cho các công ty nước ngoài của các công việc theo thứ tự giảm dần.
- Điều chỉnh các tham số và cài đặt các label/title..., sao cho hình ảnh trực quan được rõ ràng đẹp mắt.

$$T_1 = \ell_1 + 273 = 273 + 60 = 333K, T_2 = \ell_2 + 273 = 298K$$



$l + 0,5 = 1,5 \text{ mm}$

# Câu hỏi 4

**Tỉ lệ làm việc cho công ty nước ngoài của mỗi công việc (một nhân viên được xem là làm việc cho công ty nước ngoài nếu có employee\_residence khác company\_location)?**

Code bước 1:

```
df=salaries_df[['job_title','employee_residence','company_location']]  
l0=list(df['job_title'].values)  
l1=list(df['employee_residence'].values)  
l2=list(df['company_location'].values)  
job_titles=[l0[i] for i in range(len(l0)) if l1[i]!=l2[i]]  
set_job_titles=list(set(job_titles))  
job_titles_df=pd.DataFrame({'data':job_titles})  
percents=[]  
for i in set_job_titles:  
    percents.append(((job_titles_df['data'].value_counts()[i]/len(job_titles_df))*100).round(1))  
job_titles_series=pd.Series(index=set_job_titles,data=percents)  
job_titles_series=job_titles_series.sort_values(ascending=False)
```

# Câu hỏi 4

**Tỉ lệ làm việc cho công ty nước ngoài của mỗi công việc (một nhân viên được xem là làm việc cho công ty nước ngoài nếu có employee\_residence khác company\_location)?**

Hãy nhìn vào pandas.Series chứa dữ liệu cần thiết cho câu hỏi này:

Data Scientist	31.5
Data Engineer	28.1
Data Analyst	22.6
Machine Learning Engineer	8.8
Data Architect	3.1
Machine Learning Scientist	2.2
Research Scientist	1.4
Computer Vision Engineer	0.9
AI Scientist	0.8
Applied Scientist	0.4
NLP Engineer	0.2
dtype: float64	

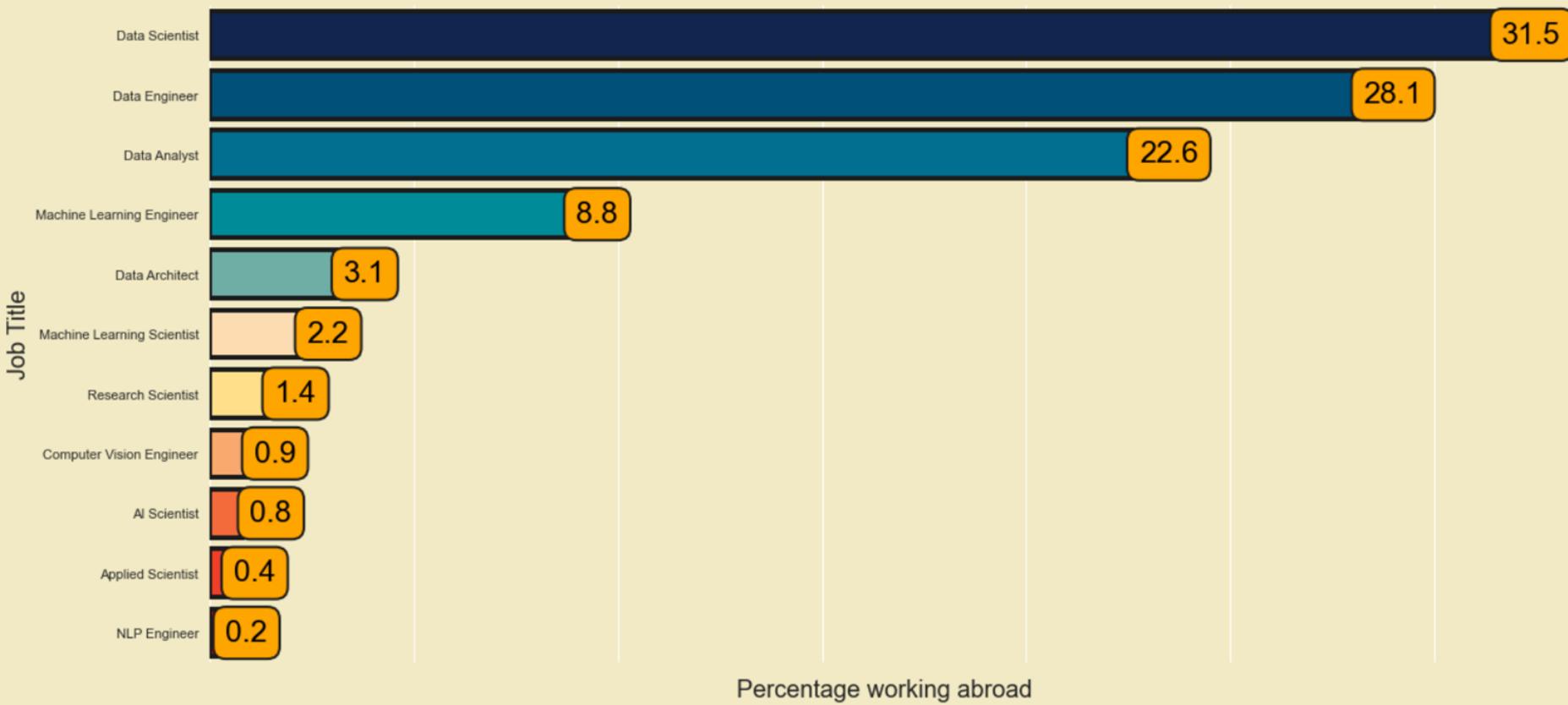
# Câu hỏi 4

**Tỉ lệ làm việc cho công ty nước ngoài của mỗi công việc (một nhân viên được xem là làm việc cho công ty nước ngoài nếu có employee\_residence khác company\_location)?**

Code bước 2:

```
plt.subplots(figsize=(20, 10))
p=sns.barplot(job_titles_series.values,job_titles_series.index,palette=palette, saturation=1,
edgecolor = "#1c1c1c", linewidth = 4)
p.axes.set_title("\nPercentage working abroad of each job\n",fontsize=25)
p.axes.set_xlabel("Percentage working abroad",fontsize=20)
p.axes.set_ylabel("Job Title",fontsize=20)
p.axes.set_xticklabels(p.get_xticklabels(),rotation = 90)
for container in p.containers:
    p.bar_label(container,label_type="edge",padding=6,size=25,color="black",rotation=0,
    bbox={"boxstyle": "round", "pad": 0.4, "facecolor": "orange", "edgecolor": "#1c1c1c",
    "linewidth" : 2, "alpha": 1})
sns.despine(left=True, bottom=True)
plt.show()
```

## Percentage working abroad of each job





$$T_1 = t_1 + 273 = 273 + 60 = 333K, T_2 = t_2 + 273 = 298K$$

$l_1 + 0,5 = 1,5 \text{ mm}$

# Câu hỏi 4

**Tỉ lệ làm việc cho công ty nước ngoài của mỗi công việc (một nhân viên được xem là làm việc cho công ty nước ngoài nếu có employee\_residence khác company\_location)?**

Nhận xét:

- Những vai trò quen thuộc luôn chiếm tỉ lệ cao nhất là: Data Scientist, Data Engineer, Data Analyst, Machine Learning Engineer.
- Để hiểu khi với sự phát triển mạnh mẽ của các ngành khoa học dữ liệu và khoa học máy tính ngày nay thì các công việc vừa nêu là những việc đòi hỏi nguồn nhân lực rất cao trong thời đại ngày nay do đó không chỉ là môi trường trong nước mà dựa vào việc đầu tư trực tiếp nước ngoài của các quốc gia đang ngày càng được đẩy mạnh thì tỉ lệ làm việc cho các công ty nước ngoài ngày càng cao.



$$T_1 = t_1 + 273 = 273 + 60 = 333K, T_2 = t_2 + 273 = 298K$$

# Câu hỏi 4

**Tỉ lệ làm việc cho công ty nước ngoài của mỗi công việc (một nhân viên được xem là làm việc cho công ty nước ngoài nếu có employee\_residence khác company\_location)?**

Nhận xét:

- Các nhà khoa học dữ liệu và khoa học máy tính rất cần thiết cho không chỉ lĩnh vực IT mà còn có vai trò quan trọng trong các vấn đề kinh doanh liên quan đến công nghệ do đó tỉ lệ cao hơn những vai trò hiếm thấy hơn như AI hay NLP Engineer.

# Câu hỏi 5

**Mức lương trung bình giữa các loại hình lao động (employment\_type) của các quy mô công ty (company\_size)?**

Ý nghĩa khi trả lời được câu hỏi: Có sự so sánh mức lương trung bình giữa các loại hình lao động của các quy mô công ty. Từ đó, đưa ra lựa chọn loại hình lao động phù hợp nhất đối với quy mô công ty mong muốn.

Tiền xử lý: Không cần thiết cho câu hỏi này.

# Câu hỏi 5

**Mức lương trung bình giữa các loại hình lao động (employment\_type) của các quy mô công ty (company\_size)?**

Phân tích dữ liệu để trả lời cho các câu hỏi:

- *Tiến hành trực quan hóa bằng barplot (thư viện seaborn) trong đó: trục hoành là quy mô công ty (company\_size), trục tung sẽ là mức lương trung bình (salary\_in\_usd) và trục hue sẽ là loại hình lao động (employment\_type)*
- *Điều chỉnh các tham số và cài đặt các label/title..., sao cho hình ảnh trực quan được rõ ràng đẹp mắt.*

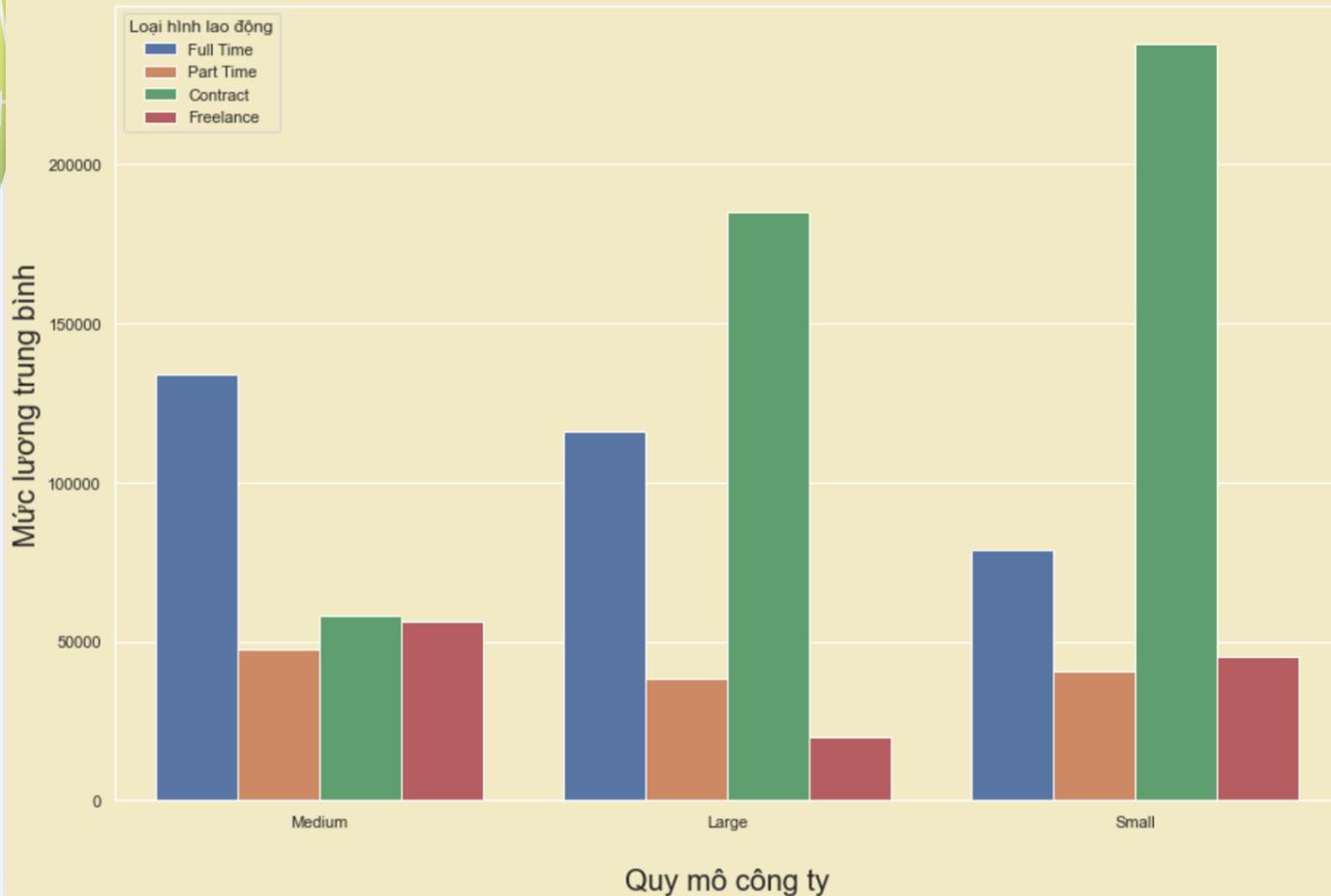
# Câu hỏi 5

**Mức lương trung bình giữa các loại hình lao động (employment\_type) của các quy mô công ty (company\_size)?**

Code:

```
plt.subplots(figsize=(15, 10))
p = sns.barplot(data=salaries_df, x='company_size', y='salary_in_usd',
hue='employment_type', ci=None)
p.set_title('\nSo sánh mức lương trung bình giữa các loại hình lao động của\nnhững quy mô công ty\n', fontsize=25)
p.set_xlabel('\nQuy mô công ty', fontsize=20)
p.set_ylabel('Mức lương trung bình', fontsize=20)
plt.legend(title='Loại hình lao động')
plt.show()
```

## So sánh mức lương trung bình giữa các loại hình lao động của các quy mô công ty



# Câu hỏi 5

**Mức lương trung bình giữa các loại hình lao động (employment\_type) của các quy mô công ty (company\_size)?**

Nhận xét:

- Loại hình lao động hợp đồng ở quy mô công ty nhỏ có mức lương trung bình cao hơn hết và trên 200 000 USD. Đây cũng là loại hình cao nhất trong quy mô công ty lớn nhưng dưới 200 000 USD.
- Loại hình lao động toàn thời gian thích hợp nhất cho quy mô công ty vừa vì loại hình này cao hơn khi trong quy mô lớn và nhỏ, cũng như cao nhất trong quy mô vừa.

# Câu hỏi 5

**Mức lương trung bình giữa các loại hình lao động (employment\_type) của các quy mô công ty (company\_size)?**

Nhận xét:

- Loại hình bán thời gian và tự do có mức lương trung bình được chi trả xấp xỉ nhau trong quy mô vừa và nhỏ; quy mô công ty vừa có phần nhỉnh hơn quy mô nhỏ đối với việc chi trả cho cả hai loại hình này.

# Câu hỏi 6

**Xu hướng của các hình thức làm việc từ xa (remote\_ratio) theo từng năm?**

Ý nghĩa khi trả lời được câu hỏi: Có cái nhìn tổng quát về sự thay đổi ở số lượng và tỉ trọng của các hình thức làm việc từ xa trong ba năm (2020, 2021, 2022). Từ đó, đánh giá được xu hướng của loại hình làm việc từ xa nói chung và các hình thức làm việc từ xa nói riêng.

Tiền xử lý: Không cần thiết cho câu hỏi này.

# Câu hỏi 6

**Xu hướng của các hình thức làm việc từ xa (remote\_ratio) theo từng năm?**

Phân tích dữ liệu để trả lời cho các câu hỏi:

Bước 1:

- Tạo dataframe 'remote\_per\_year' gồm:
- 2 index level: 'work\_year' (lưu 3 năm: 2020, 2021, 2022) và 'remote\_ratio' (lưu các hình thức làm việc từ xa: Fully Remote, Partially Remote, On-site);
- 2 thuộc tính : 'number' (lưu số lượng tương ứng) và percentage (lưu tỉ lệ phần trăm tương ứng).

# Câu hỏi 6

**Xu hướng của các hình thức làm việc từ xa (remote\_ratio) theo từng năm?**

Phân tích dữ liệu để trả lời cho các câu hỏi:

Bước 2:

- Vẽ linechart để minh họa cho sự thay đổi ở số lượng của các hình thức làm việc từ xa theo từng năm.
- Vẽ từng piechart để minh họa cho sự thay đổi ở tỉ trọng của các hình thức làm việc từ xa theo từng năm tương ứng.
- Điều chỉnh các tham số và cài đặt các label/title..., sao cho hình ảnh trực quan được rõ ràng đẹp mắt.

# Câu hỏi 6

**Xu hướng của các hình thức làm việc từ xa (remote\_ratio) theo từng năm?**

Code bước 1:

```
remote_per_year = pd.DataFrame(salaries_df[['work_year',  
'remote_ratio']].groupby('work_year').value_counts())  
remote_per_year['percentage'] = remote_per_year.groupby(level=0).apply(lambda x:  
100 * x / float(x.sum()))  
remote_per_year.columns = ['number', 'percentage']
```

# Câu hỏi 6

**Xu hướng của các hình thức làm việc từ xa (remote\_ratio) theo từng năm?**

Hãy nhìn vào dataframe chứa dữ liệu cần thiết cho câu hỏi này:

work_year	remote_ratio	number	percentage
2020	Fully Remote	39	52.000000
	Partially Remote	21	28.000000
	On-site	15	20.000000
2021	Fully Remote	119	52.654867
	Partially Remote	73	32.300885
	On-site	34	15.044248
2022	Fully Remote	662	59.001783
	On-site	413	36.809269
	Partially Remote	47	4.188948

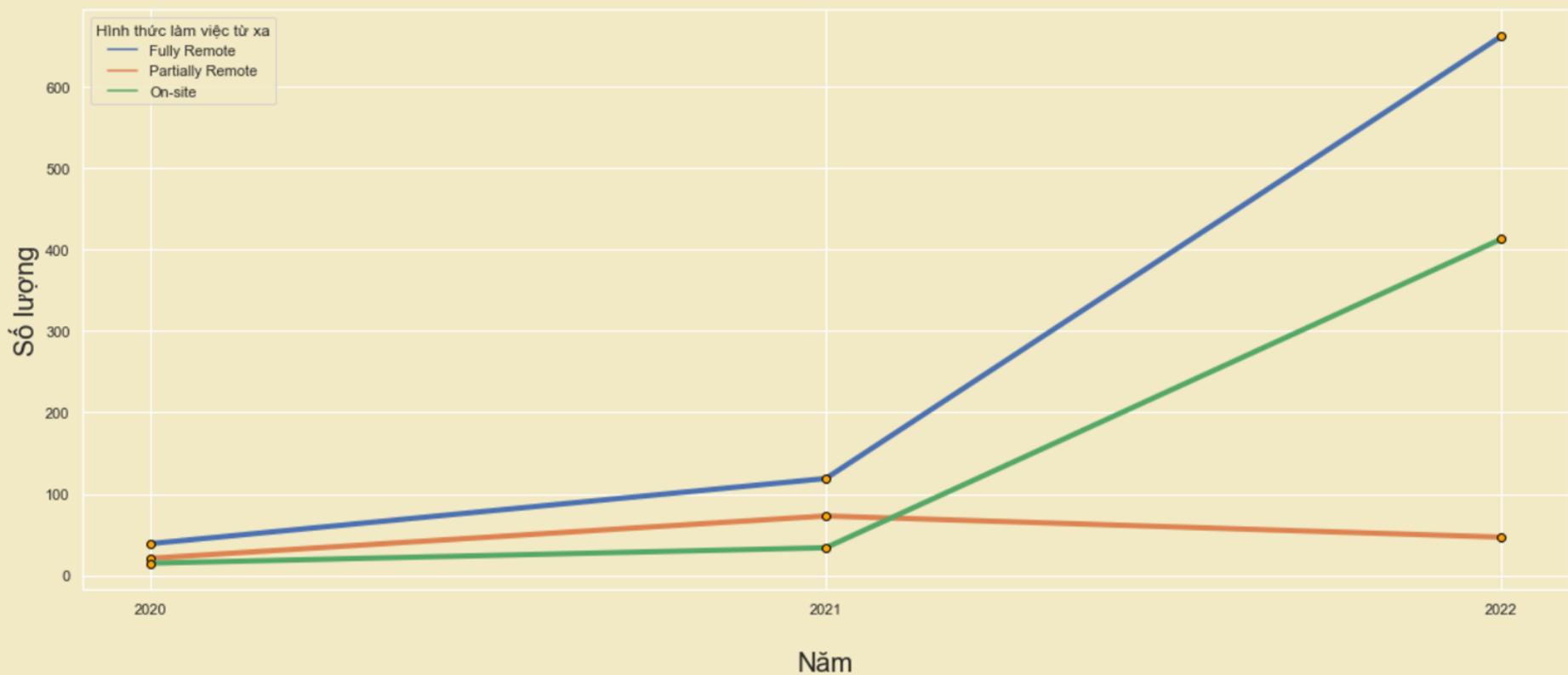
# Câu hỏi 6

**Xu hướng của các hình thức làm việc từ xa (remote\_ratio) theo từng năm?**

Code bước 2:

```
plt.subplots(figsize=(20, 8))
p = sns.lineplot(data=remote_per_year, x='work_year', y='number',
hue='remote_ratio', marker="o", linewidth=4, markersize=6, markerfacecolor="orange", m
arkeredgecolor="black", markeredgewidth=1)
p.set_title('\nSố lượng của các hình thức làm việc từ xa theo từng năm\n',
fontsize=25)
p.set_xlabel('\nNăm', fontsize=20)
p.set_ylabel('Số lượng', fontsize=20)
p.set_xticks(remote_per_year.index.get_level_values(0).unique())
plt.legend(title='Hình thức làm việc từ xa')
plt.show()
```

## Số lượng của các hình thức làm việc từ xa theo từng năm



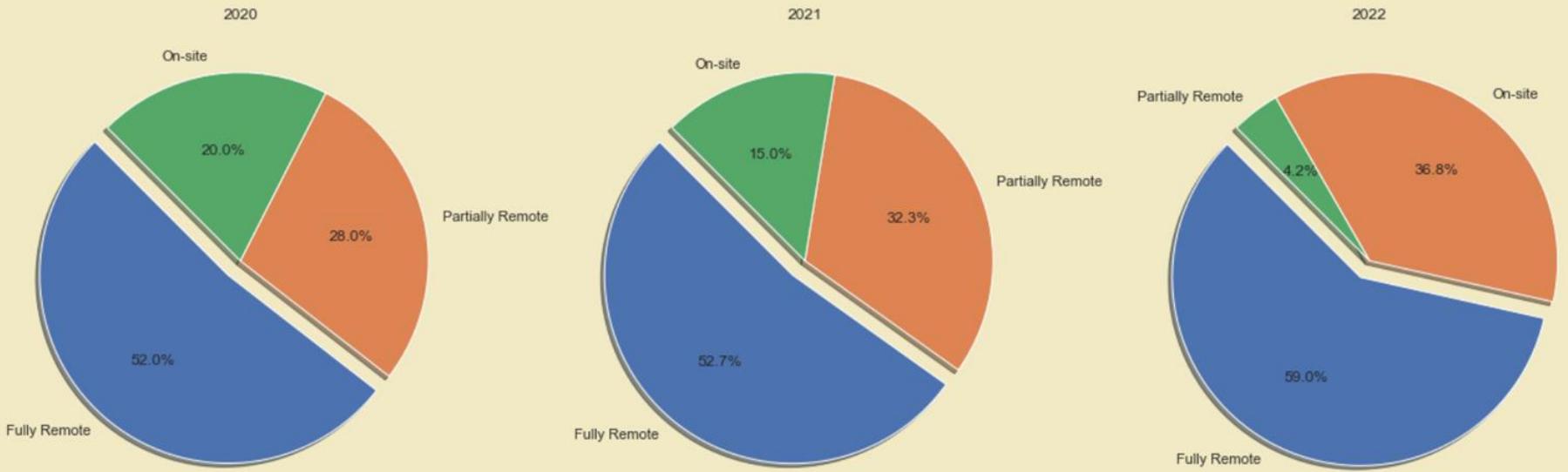
# Câu hỏi 6

**Xu hướng của các hình thức làm việc từ xa (remote\_ratio) theo từng năm?**

Tiếp tục code bước 2:

```
fig, ax = plt.subplots(1,  
len(remote_per_year.index.get_level_values(0).unique()), figsize=(22, 8))  
fig.suptitle('\nTỉ trọng của các hình thức làm việc từ xa theo từng năm',  
fontsize=25)  
print(f'Chúng ta hãy nhìn vào sự thay đổi ở tỉ trọng của các hình thức làm việc  
từ xa theo từng năm: ')  
explode = (0.1, 0, 0)  
for i in range(len(remote_per_year.index.get_level_values(0).unique())):  
    pie_df = remote_per_year.xs(2020 + i, level=0, axis=0, drop_level=False)  
    ax[i].pie(pie_df['percentage'].values,  
labels=pie_df.index.get_level_values(1), explode=explode, autopct='%.1f%%',  
shadow=True, startangle=135)  
    ax[i].set_title(2020 + i)
```

## Tỉ trọng của các hình thức làm việc từ xa theo từng năm



# Câu hỏi 6

**Xu hướng của các hình thức làm việc từ xa (remote\_ratio) theo từng năm?**

Nhận xét:

1. Dựa trên sự thay đổi ở số lượng của các hình thức làm việc từ xa theo từng năm:

- *Hình thức làm việc từ xa toàn phần và trực tiếp đều tăng nhẹ từ 2020 đến 2021 và tăng mạnh từ 2021 đến 2022; hình thức từ xa toàn phần có phần tăng nhanh hơn hình thức trực tiếp.*
- *Hình thức làm việc từ xa bán phần giữ ở mức ổn định (không biến động mạnh) và bị hình thức trực tiếp vượt mặt nhanh từ sau năm 2021.*

# Câu hỏi 6

**Xu hướng của các hình thức làm việc từ xa (remote\_ratio) theo từng năm?**

Nhận xét:

2. Dựa trên sự thay đổi ở số lượng của các hình thức làm việc từ xa theo từng năm:

- *Hình thức làm việc từ xa toàn phần luôn chiếm tỉ trọng lớn (hơn 50%) và tăng đều.*
- *Hình thức làm việc trực tiếp luôn chiếm tỉ trọng lớn hơn 25% và tăng đều.*
- *Hình thức làm việc từ xa bán phần giảm mạnh.*

# Câu hỏi 6

**Xu hướng của các hình thức làm việc từ xa (remote\_ratio) theo từng năm?**

Nhận xét:

- Nếu ta giả định tình hình đại dịch Covid-19 ảnh hưởng đến các sự thay đổi này, các nhận xét trên có thể lý giải và kết luận như sau:
- Số lượng của hình thức làm việc trực tiếp giữa 2020 và 2021 bị ảnh hưởng bởi đại dịch nên ít hơn 2 hình thức còn lại, nhưng lại tăng mạnh sau 2021 do đã bình thường hóa Covid-19.
- Số lượng và tỉ trọng của từ xa toàn phần luôn tăng cho thấy đây là hình thức làm việc được ưa chuộng trong những năm sắp tới.
- Hình thức làm việc từ xa bán phần dần bị đào thải do sự giảm mạnh trong tỉ trọng theo từng năm.

# Câu hỏi 7

**Hãy so sánh để tìm ra mức lương trung bình giữa các mức độ kinh nghiệm (experience level) qua các năm?**

Ý nghĩa khi trả lời được câu hỏi: Có cái nhìn chung về sự thay đổi ở mức lương của các mức kinh nghiệm trong ba năm (2020, 2021, 2022). Từ đó, đánh giá được sự ảnh hưởng của kinh nghiệm làm việc đối với mức lương.

Tiền xử lý: Không cần thiết cho câu hỏi này.

# Câu hỏi 7

**Hãy so sánh để tìm ra mức lương trung bình giữa các mức độ kinh nghiệm (experience level) qua các năm?**

Phân tích dữ liệu để trả lời cho các câu hỏi:

Bước 1: Tạo dataframe 'year\_salary' gồm:

- cột 'work\_year' lưu 3 năm: 2020, 2021, 2022
- cột 'experience\_level' lưu các mức kinh nghiệm: 'Entry-level / Junior', 'Executive-level / Director', 'Mid-level / Intermediate', 'Senior-level / Expert'
- cột 'salary\_in\_usd' lưu mức lương trung bình của các mức kinh nghiệm theo từng năm.

# Câu hỏi 7

**Hãy so sánh để tìm ra mức lương trung bình giữa các mức độ kinh nghiệm (experience level) qua các năm?**

Phân tích dữ liệu để trả lời cho các câu hỏi:

Bước 2:

- Vẽ linechart để minh họa cho sự biến đổi ở mức lương trung bình của các mức kinh nghiệm theo từng năm.
- Điều chỉnh các tham số và cài đặt các label/title..., sao cho hình ảnh trực quan được rõ ràng đẹp mắt.

# Câu hỏi 7

**Hãy so sánh để tìm ra mức lương trung bình giữa các mức độ kinh nghiệm (experience level) qua các năm?**

Code bước 1:

```
year_salary=salaries_df.groupby(["work_year","experience_level"])[  
    "salary_in_usd"].mean()  
year_salary=year_salary.reset_index()
```

# Câu hỏi 7

**Hãy so sánh để tìm ra mức lương trung bình giữa các mức độ kinh nghiệm (experience level) qua các năm?**

Hãy nhìn vào dataframe chứa dữ liệu cần thiết cho câu hỏi này:

	work_year	experience_level	salary_in_usd
0	2020	Entry-level / Junior	57511.608696
1	2020	Executive-level / Director	202416.500000
2	2020	Mid-level / Intermediate	85950.062500
3	2020	Senior-level / Expert	137240.500000
4	2021	Entry-level / Junior	54385.537037
5	2021	Executive-level / Director	186128.000000
6	2021	Mid-level / Intermediate	81323.977778
7	2021	Senior-level / Expert	125557.458333
8	2022	Entry-level / Junior	65309.410256
9	2022	Executive-level / Director	183878.406250
10	2022	Mid-level / Intermediate	95282.556522
11	2022	Senior-level / Expert	148454.282609

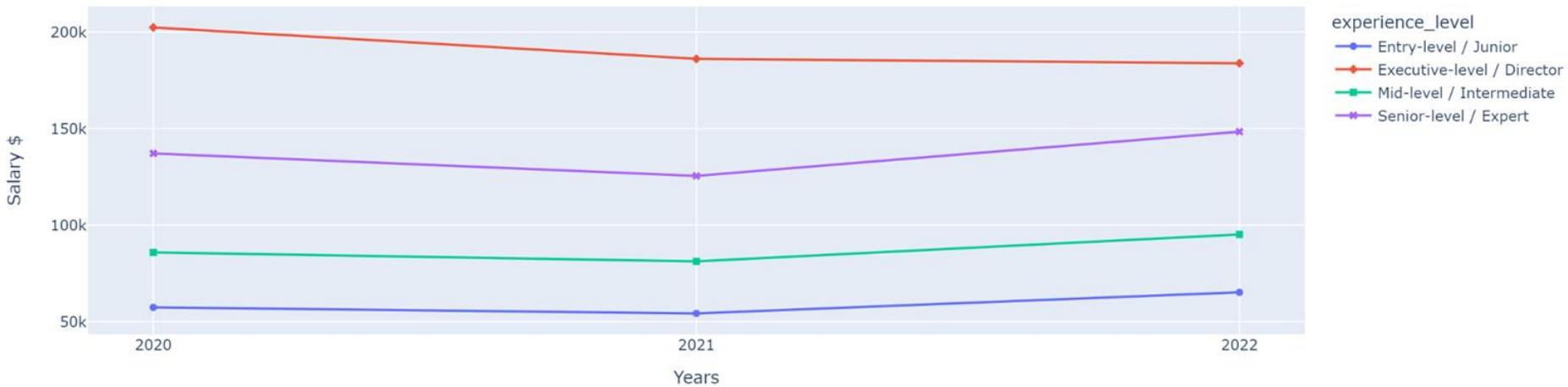
# Câu hỏi 7

**Hãy so sánh để tìm ra mức lương trung bình giữa các mức độ kinh nghiệm (experience level) qua các năm?**

Code bước 2:

```
fig=px.line(year_salary,x="work_year",y="salary_in_usd",color="experience_level", symbol="experience_level", title='Lương trung bình của các mức kinh nghiệm theo từng năm')  
fig.update_layout(yaxis_title="Salary$",xaxis_title="Years",xaxis=dict(tickmode='array',tickvals=[2020, 2021, 2022]))
```

### Lương trung bình của các mức kinh nghiệm theo từng năm



# Câu hỏi 7

**Hãy so sánh để tìm ra mức lương trung bình giữa các mức độ kinh nghiệm (experience level) qua các năm?**

Nhận xét:

Mức lương trung bình của từng level có sự chênh lệch đáng kể.

- Level 'Executive-level / Director' có mức lương trung bình cao nhất, dao động từ 180.000 đến 200.000 USD mỗi năm. Mức lương giám đốc giảm dần từ 2020 đến 2022.
- Level 'Entry-level / Junior' có mức lương trung bình thấp nhất, dao động từ 54.000 đến 65.000 USD. Giảm nhẹ từ 2020 đến 2021, tăng mạnh từ 2021 đến 2022.

# Câu hỏi 7

**Hãy so sánh để tìm ra mức lương trung bình giữa các mức độ kinh nghiệm (experience level) qua các năm?**

Nhận xét:

- Hai level 'Mid-level / Intermediate' và 'Senior-level / Expert' có mức lương trung bình ở mức trung bình, dao động từ 80.000 đến 95.000 USD mỗi năm đối với level 'Mid-level / Intermediate' và 125.000 đến 150.000 đối với level 'Senior-level / Expert'. Mức lương của cả hai level này cũng giảm nhẹ vào năm 2021 và tăng mạnh vào năm 2022.

# Câu hỏi 8

**Đâu là top 10 vai trò có mức lương trung bình cao nhất?**

Ý nghĩa khi trả lời được câu hỏi: Câu hỏi này giúp bạn hiểu được các vai trò có mức lương cao nhất, từ đó có thể đưa ra các đánh giá xu hướng việc làm trong tương lai.

Tiền xử lý: Không cần thiết cho câu hỏi này.

# Câu hỏi 8

**Đâu là top 10 vai trò có mức lương trung bình cao nhất?**

Phân tích dữ liệu để trả lời cho các câu hỏi:

Bước 1: Tạo dataframe 'role\_salary' gồm:

- cột 'job\_title' lưu các tên công việc (vai trò).
- cột 'salary\_in\_usd' lưu mức lương cao nhất của từng vai trò.

Bước 2:

- Vẽ barchart để minh họa cho mức lương cao nhất của các vai trò.
- Điều chỉnh các tham số và cài đặt các label/title..., sao cho hình ảnh trực quan được rõ ràng đẹp mắt.

# Câu hỏi 8

**Đâu là top 10 vai trò có mức lương trung bình cao nhất?**

Code bước 1:

```
role_salary =  
salaries_df.groupby('job_title',as_index=False)[ 'salary_in_usd' ]  
.mean()  
role_salary =  
role_salary.sort_values(by='salary_in_usd',ascending=False)
```

# Câu hỏi 8

**Đâu là top 10 vai trò có mức lương trung bình cao nhất?**

Hãy nhìn vào dataframe chứa dữ liệu cần thiết cho câu hỏi này:

	job_title	salary_in_usd
1	Applied Scientist	179600.000000
4	Data Architect	163759.477273
6	Data Scientist	134376.689732
5	Data Engineer	129685.280000
8	Machine Learning Scientist	128871.064516
7	Machine Learning Engineer	127952.096000
10	Research Scientist	108845.950000
3	Data Analyst	102503.697819
0	AI Scientist	78964.333333
2	Computer Vision Engineer	47236.000000
9	NLP Engineer	33686.000000

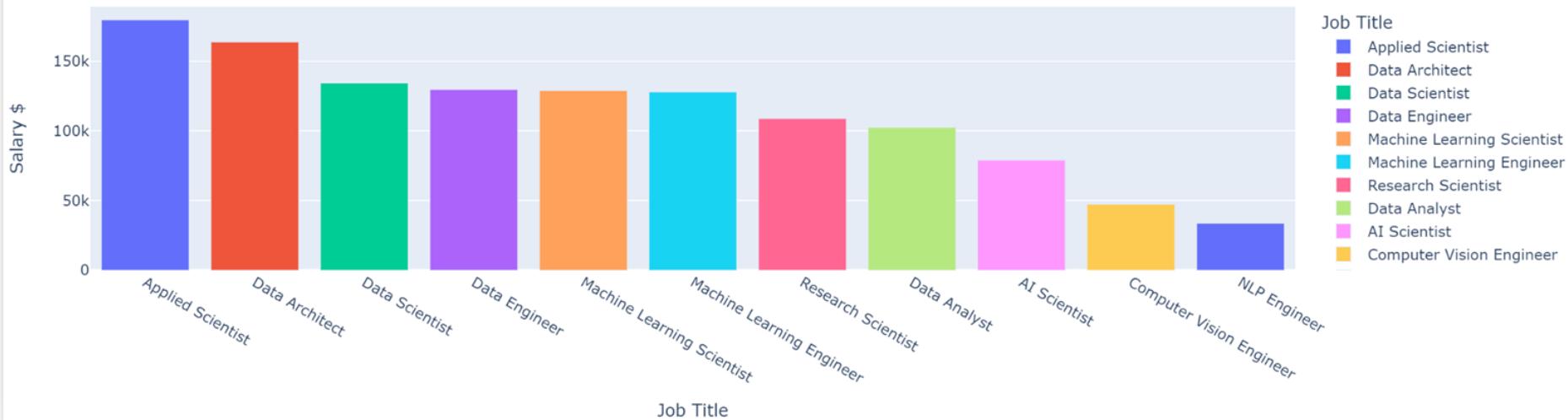
# Câu hỏi 8

**Đâu là top 10 vai trò có mức lương trung bình cao nhất?**

Code bước 2:

```
fig=px.bar(role_salary,  
x='job_title',y='salary_in_usd',color='job_title',  
    labels={'job_title':'Job Title','salary_in_usd':'Salary $'},  
    title='Top 10 Vai trò được trả lương cao nhất trong Khoa học  
dữ liệu')  
fig.show()
```

## Top 10 Vai trò được trả lương cao nhất trong Khoa học dữ liệu



# Câu hỏi 8

**Đâu là top 10 vai trò có mức lương trung bình cao nhất?**

Nhận xét:

- Đứng đầu bảng xếp hạng là vai trò "Applied Scientist" với mức lương trung bình 180.000 USD mỗi năm. Khoa học ứng dụng là ngành khoa học sử dụng phương pháp khoa học và kiến thức thu được thông qua các kết luận từ phương pháp để đạt được các mục tiêu thực tiễn. Có thể nói đây là một trong những ngành học có mức lương cao nhất hiện nay.

# Câu hỏi 8

**Đâu là top 10 vai trò có mức lương trung bình cao nhất?**

Nhận xét:

- Tiếp đến các ngành về Data: Data Architect, Data Scientist và Data Engineer đều có mức lương trung bình khá cao từ 130.000 đến 160.000 USD mỗi năm. Đây là những ngành học có xu hướng phát triển rất mạnh trong thời gian tới.

# Resources

<https://www.datacamp.com/cheat-sheet/jupyter-notebook-cheat-sheet>

<https://www.datacamp.com/cheat-sheet/pandas-cheat-sheet-for-data-science-in-python>

<https://slidesgo.com/theme/computer-science-mathematics-major-for-college-applied-mathematics>

$$\iiint x^2 dx dy dz =$$

$$V: z = 10(x+3y), x+y=1$$

$$x=0, y=0, z=0$$

$$= \int_0^1 \int_0^{1-x} \int_{-10(x+3y)}^5 dx dy dz$$

$$\int_0^{1/\sqrt{2}} dy \int_{-\infty}^{\arcsin y} dx$$

$$2\sqrt{y^2 - x^2}$$



$$x = 2y^2 + 3, x = 5$$

$$z = 1 + \sqrt{9x^2 + 4y^2}$$

$$z = 4 + \sqrt{9x^2 + 4y^2}$$

$$V = \int_{-1}^1 dy \int_{2y^2+3}^5 dx \int_{1+\sqrt{9x^2+4y^2}}^{4+\sqrt{9x^2+4y^2}} dz$$



# Thanks!

Do you have any questions?