Labor market analysis
Data Science major
2020 - 2023

Table of contents

1 Introduction to datasets

2 Python processing

3 Power BI analysis

4 Conclusion & Predictions

Introduction to datasets

Introduction to datasets

Data Science Salaries 2023:

Salaries of different fields within the data science industry

- **work_year**: Fiscal year.
- **experience_level**: Level of relevant experience
- **working_type**: Employment type for the role
- **job_title**: Position worked during the year
- **salary**: Total gross salary paid
- **salary_currency**: Currency of salary paid (ISO 4217 currency code)
- **salary_in_usd**: Salary in USD
- **employee_residence**: The employee's primary country of residence during the year of employment (ISO 3166 country code)
- **remote_ratio**: Total amount of work performed remotely
- **company_location**: Country of the employer's main office or contracting branch
- **company_size**: Average number of people working for the company

	work_year	experience_level	employment_type	job_title	sa
0	2023	SE	FT	Principal Data Scientist	8
1	2023	MI	CT	ML Engineer	3
2	2023	MI	CT	ML Engineer	2
3	2023	SE	FT	Data Scientist	17
4	2023	SE	FT	Data Scientist	12
...	...	...	...	...	
3750	2020	SE	FT	Data Scientist	41
3751	2021	MI	FT	Principal Data Scientist	15
3752	2020	EN	FT	Data Scientist	10
3753	2020	EN	CT	Business Data Analyst	10

salary	salary_currency	salary_in_usd	employee_residence	remote_ratio	company
80000	EUR	85847	ES	100	
30000	USD	30000	US	100	
25500	USD	25500	US	100	
175000	USD	175000	CA	100	
120000	USD	120000	CA	100	
...	...	...	...	...	
412000	USD	412000	US	100	
151000	USD	151000	US	100	
105000	USD	105000	US	100	
100000	USD	100000	US	100	
7000000	INR	94665	IN	50	

Python processing

Data Checking
Data Cleansing
Data Transformation

Python processing

1. Upload Dataset
2. Data evaluation

```
# Đọc dữ liệu từ ds_salaries và lưu vào DataFrame
import pandas as pd
from google.colab import files
uploaded = files.upload()

 sl=pd.read_csv('ds_salaries.csv')
```

```
sl.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3755 entries, 0 to 3754
Data columns (total 11 columns):
 #   Column              Non-Null Count  Dtype 
---  ------              --------------  ----- 
 0   work_year           3755 non-null   int64 
 1   experience_level    3755 non-null   object
 2   employment_type     3755 non-null   object
 3   job_title           3755 non-null   object
 4   salary              3755 non-null   int64 
 5   salary_currency     3755 non-null   object
 6   salary_in_usd       3755 non-null   int64 
 7   employee_residence  3755 non-null   object
 8   remote_ratio        3755 non-null   int64 
 9   company_location    3755 non-null   object
 10  company_size        3755 non-null   object
dtypes: int64(4), object(7)
memory usage: 322.8+ KB
```

```
# Kiểm tra giá trị Null
sl.isnull().sum()
```

```
work_year             0
experience_level      0
employment_type       0
job_title             0
salary                0
salary_currency       0
salary_in_usd         0
employee_residence    0
remote_ratio          0
company_location      0
company_size          0
dtype: int64
```

2. Data evaluation

```
sl.head()
```

	work_year	experience_level	employment_type	job_title	salary	salary_currency	salary_in_usd	employee_residence	remote_ratio	company_location	company_size
0	2023	SE	FT	Principal Data Scientist	80000	EUR	85847	ES	100	ES	L
1	2023	MI	CT	ML Engineer	30000	USD	30000	US	100	US	S
2	2023	MI	CT	ML Engineer	25500	USD	25500	US	100	US	S
3	2023	SE	FT	Data Scientist	175000	USD	175000	CA	100	CA	M
4	2023	SE	FT	Data Scientist	120000	USD	120000	CA	100	CA	M

```
sl[sl.duplicated()]
```

	work_year	experience_level	employment_type	job_title	salary	salary_currency	salary_in_usd	employee_residence	remote_ratio	company_location	company_size
115	2023	SE	FT	Data Scientist	150000	USD	150000	US	0	US	M
123	2023	SE	FT	Analytics Engineer	289800	USD	289800	US	0	US	M
153	2023	MI	FT	Data Engineer	100000	USD	100000	US	100	US	M
154	2023	MI	FT	Data Engineer	70000	USD	70000	US	100	US	M
160	2023	SE	FT	Data Engineer	115000	USD	115000	US	0	US	M
...	...	...	...	...	...	...	...	...	...	...	...
3439	2022	MI	FT	Data Scientist	78000	USD	78000	US	100	US	M
3440	2022	SE	FT	Data Engineer	135000	USD	135000	US	100	US	M
3441	2022	SE	FT	Data Engineer	115000	USD	115000	US	100	US	M
3586	2021	MI	FT	Data Engineer	200000	USD	200000	US	100	US	L
3709	2021	MI	FT	Data Scientist	76760	EUR	90734	DE	50	DE	L

1171 rows × 11 columns

Python processing

3. Create a primary key for the table

```
#Tạo khóa chính cho bảng
sl.insert(0, 'Job_ID', range(1, len(sl) + 1))
sl
```

	Job_ID	work_year	experience_level	employment_type	job_title	salary	salary_c
0	1	2023	SE	FT	Principal Data Scientist	80000	
1	2	2023	MI	CT	ML Engineer	30000	
2	3	2023	MI	CT	ML Engineer	25500	
3	4	2023	SE	FT	Data Scientist	175000	
4	5	2023	SE	FT	Data Scientist	120000	
...	...	...	...	...	...	...	
3750	3751	2020	SE	FT	Data Scientist	412000	
3751	3752	2021	MI	FT	Principal Data Scientist	151000	
3752	3753	2020	EN	FT	Data Scientist	105000	
3753	3754	2020	EN	CT	Business Data Analyst	100000	
3754	3755	2021	SE	FT	Data Science Manager	7000000	

3755 rows × 12 columns

Python processing

4. Add working type column according to working type
5. Add column salary_ranking according to salary level

```
#Thêm cột working type theo hình thức làm việc
sl['Working_Type'] = sl['remote_ratio'].apply(lambda x: 'Remote'
                                                if x == 100 else
                                                 ('Hybrid' if x == 50 else 'On_Site'))
```

```
#Thêm cột salary_ranking
sl['salary_ranking'] = pd.cut(sl['salary_in_usd'],
                              bins=[-float('inf'), 79000, 99000, 165000, 207000, float('inf')],
                              labels=['<79000', '79000<x<99000', '99000<x<165000', '165000<x<207000', '>207000'])
sl
```

te_ratio	company_location	company_size	Working_Type	salary_ranking
100	ES	L	Remote	79000<x<99000
100	US	S	Remote	<79000
100	US	S	Remote	<79000
100	CA	M	Remote	165000<x<207000
100	CA	M	Remote	99000<x<165000
...	...	...	...	...
100	US	L	Remote	>207000

Python processing

6. Create sub-tables using groupby

```
#Đếm số công việc theo mức lương
count_salary_ranking = sl.groupby('salary_ranking').size().reset_index(name='count')
count_salary_ranking

#Đếm số công việc theo hình thức làm việc
count_working_type = sl.groupby('Working_Type').size().reset_index(name='count')
count_working_type
```

```
#Tính trung bình lương theo hình thức làm việc
avg_salary_by_working_type = sl.groupby('Working_Type')['salary_in_usd'].mean().reset_index(name='avg_salary')
avg_salary_by_working_type

#Tính lương trung bình theo experience_level
avg_salary_by_experience_level = sl.groupby('experience_level')['salary_in_usd'].mean().reset_index()
avg_salary_by_experience_level.columns = ['experience_level', 'avg_salary_in_usd']
avg_salary_by_experience_level = avg_salary_by_experience_level.sort_values('avg_salary_in_usd', ascending=False)
avg_salary_by_experience_level
```

```
#Tính số lượng job theo residence
number_of_job_by_residence = sl['employee_residence'].value_counts().reset_index()
number_of_job_by_residence.columns = ['employee_residence', 'number_of_jobs']
number_of_job_by_residence

#Tính trung bình lương theo company_size
salary_by_company_size = sl.groupby('company_size')['salary_in_usd'].mean().reset_index()
salary_by_company_size.columns = ['company_size', 'avg_salary_in_usd']
salary_by_company_size
```

```
#Tính tổng đơn tuyển theo năm
number_of_job_per_year = sl.groupby('work_year').size().reset_index(name='number_of_jobs')
number_of_job_per_year = number_of_job_per_year.sort_values('work_year')
number_of_job_per_year
```

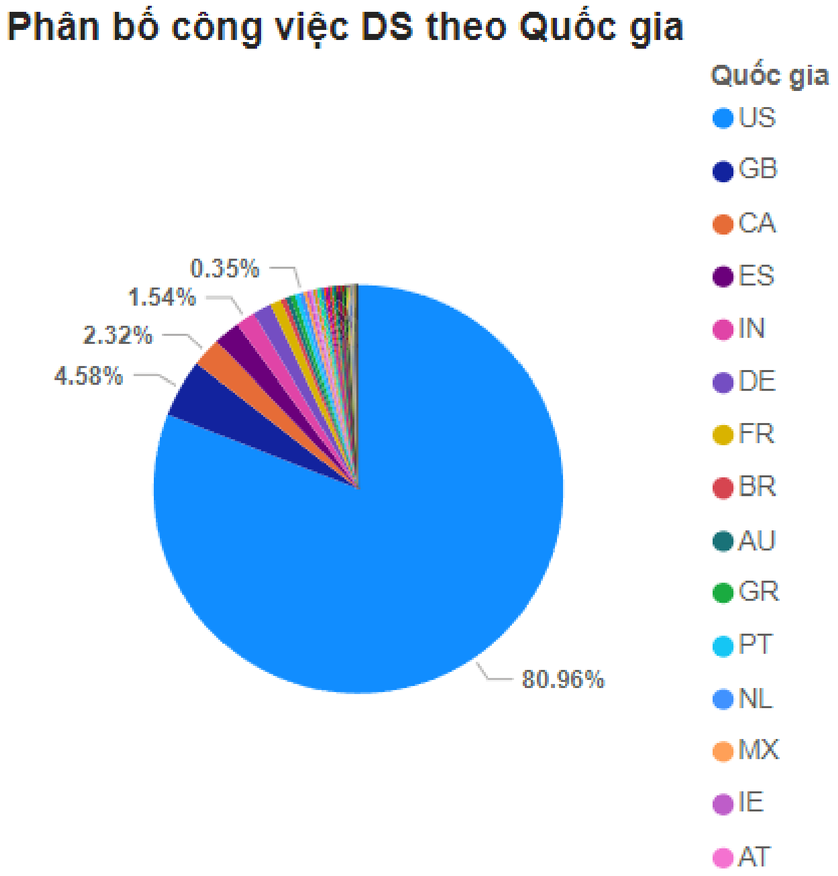
Python processing

7. Extract data

```
#Trích xuất dữ liệu
sl.to_csv('final_sl.csv')
files.download('final_sl.csv')

avg_salary_by_experience_level.to_csv('avg_salary_by_experience_level.csv')
files.download('avg_salary_by_experience_level.csv')

number_of_job_by_residence.to_csv('number_of_job_by_residence.csv')
files.download('number_of_job_by_residence.csv')

salary_by_company_size.to_csv('salary_by_company_size.csv')
files.download('salary_by_company_size.csv')

number_of_job_per_year.to_csv('number_of_job_per_year.csv')
files.download('number_of_job_per_year.csv')

avg_salary_by_job('avg_salary_by_job.csv')
files.download('avg_salary_by_job.csv')

highest_job_salary_by_location('highest_job_salary_by_location.csv')
files.download('highest_job_salary_by_location.csv')

count_salary_ranking('count_salary_ranking.csv')
files.download('count_salary_ranking.csv')

count_working_type('count_working_type.csv')
files.download('count_working_type.csv')

avg_salary_by_working_type('avg_salary_by_working_type.csv')
files.download('avg_salary_by_working_type.csv')
```

Power BI analysis

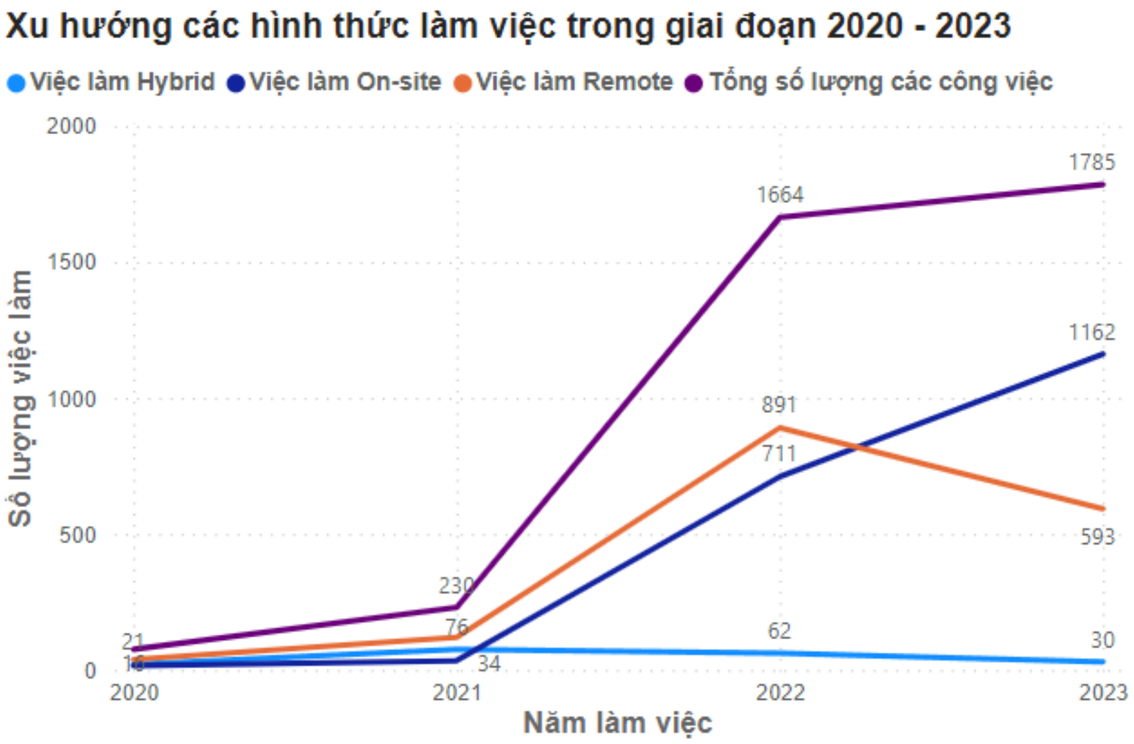
By Country



The United States is currently a fertile destination for Data Science personnel with 3,040 jobs, accounting for 80.96% of the total market workforce.

Other countries with high recruitment demand after the United States are the United Kingdom, Canada, Spain and India.

By Working type

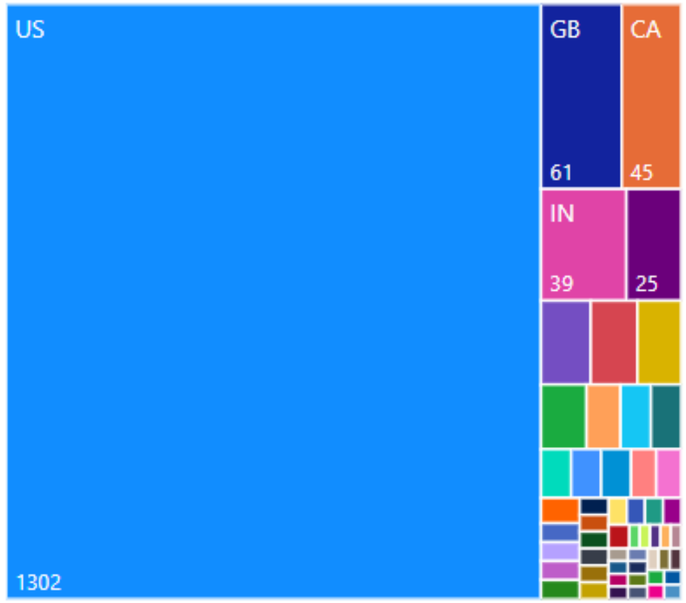


- **During 2020 - 2021, Data Science jobs are available in very limited numbers. The difference between the two years is not high, showing that companies have not yet fully recovered and employers are not ready to recruit more employees.**
- **By 2022, the growth rate will reach nearly 900% and increase by 7% by 2023. By this stage, companies have recovered and are on the path to growth again, the need for human resources is increasing.**

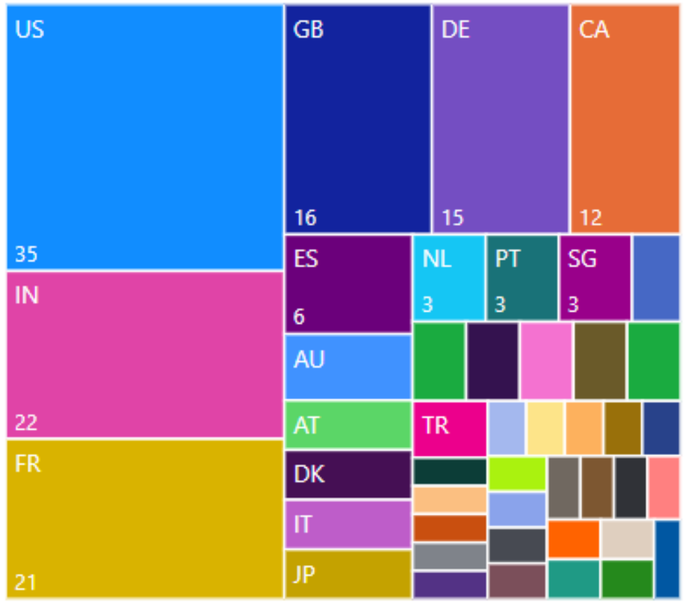
- With Remote work: accounting for the highest number in the period 2020 - 2021. This is inevitable because during the Covid period, most employees have to work from home. By 2022, the number of jobs will also increase to 891, but will decrease deeply by 34% to 593 in 2023.
- With On-site jobs: reaching the lowest number during the covid period 2020 - 2021. Since then, the number of jobs has continuously increased as people returned to the office. 2022 sets a 96% increase. The 39% increase in 2023 may come from a decline in Remote jobs during the same period.
- Hybrid employment has not changed much in 4 years. In the period 2020 - 2021, the number of jobs is the second largest after Remote.

By Country of residence

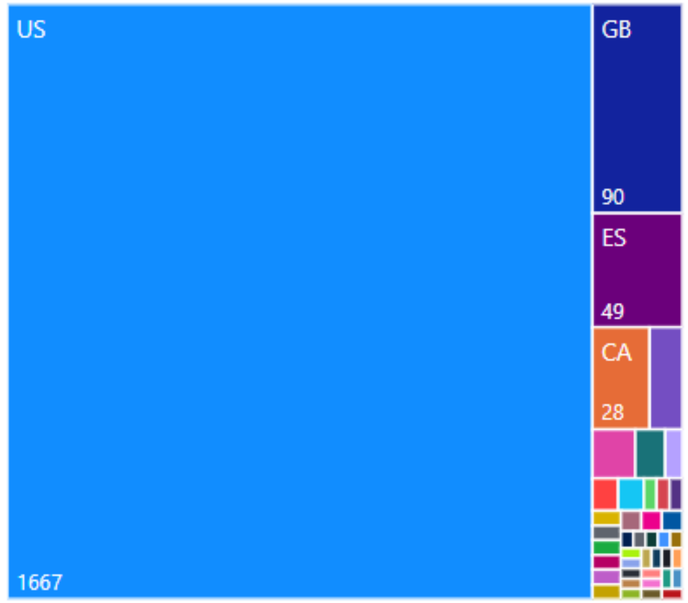
Phân bổ khu vực sinh sống của nhân sự Remote



Phân bổ khu vực sinh sống của nhân sự Hybrid

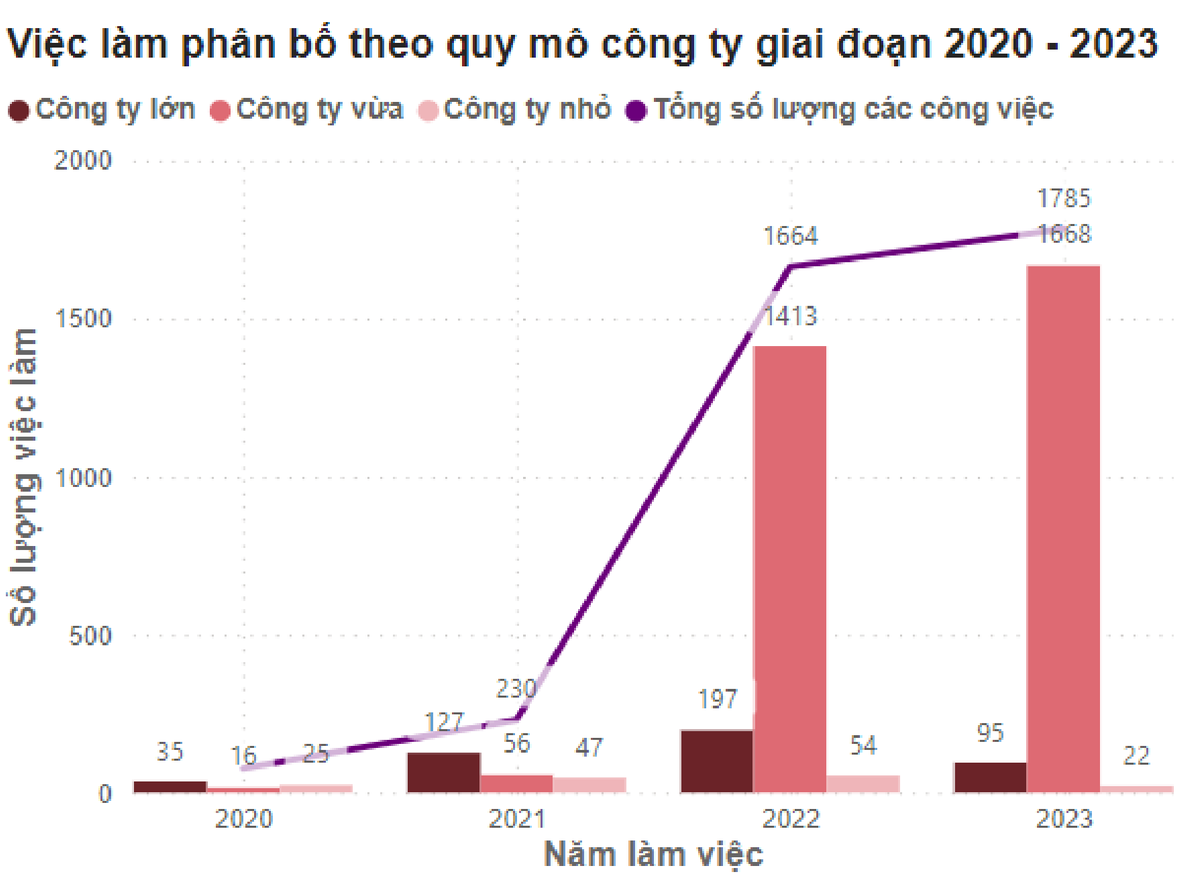


Phân bổ khu vực sinh sống của nhân sự On-site



- Most of the employees working Remotely and On-site live in the United States.
- In the countries of residence, the number of Hybrid employees is distributed quite evenly, with the largest number in groups of countries such as the United States, United Kingdom, Germany, India, Canada, and France.
- The Data Science workforce mostly lives in the United States, working Remotely and On-site with a ratio of 44 - 56.

By Company size



- **In the period 2020 - 2021, large companies are the companies maintaining the most Data Science jobs.**
- **Going into the 2022 - 2023 period, the companies with the most Data Science jobs are medium-sized companies, more than 7 times the total number of jobs in the other two sectors. The difference will be even higher in 2023, from 7 times to more than 15 times.**

- For large companies: maintain the most jobs in the covid period 2020 - 2021. The number of jobs follows the general growth momentum in 2022, but decreases by more than 50% in 2023.
- For medium companies: although the number is the smallest in the period 2020, the number of jobs increases in line with the market momentum, being the main driver of growth in the number of jobs in the period 2022 - 2023, with an increase of 96% in 2022 and 15% in 2023.
- Small companies: increase steadily according to the general growth momentum until 2022, then decline by more than 50% in 2023.

CONCLUSION

By Country

The United States is the destination with the most jobs for Data Science personnel

By Working type

After Covid and when the economy tends to recover, On-site working form still has the highest trend. Remote working has and will likely decrease over time as society fully normalizes

By Country of residence

The Data Science workforce mostly lives in the United States, working Remotely and On-site

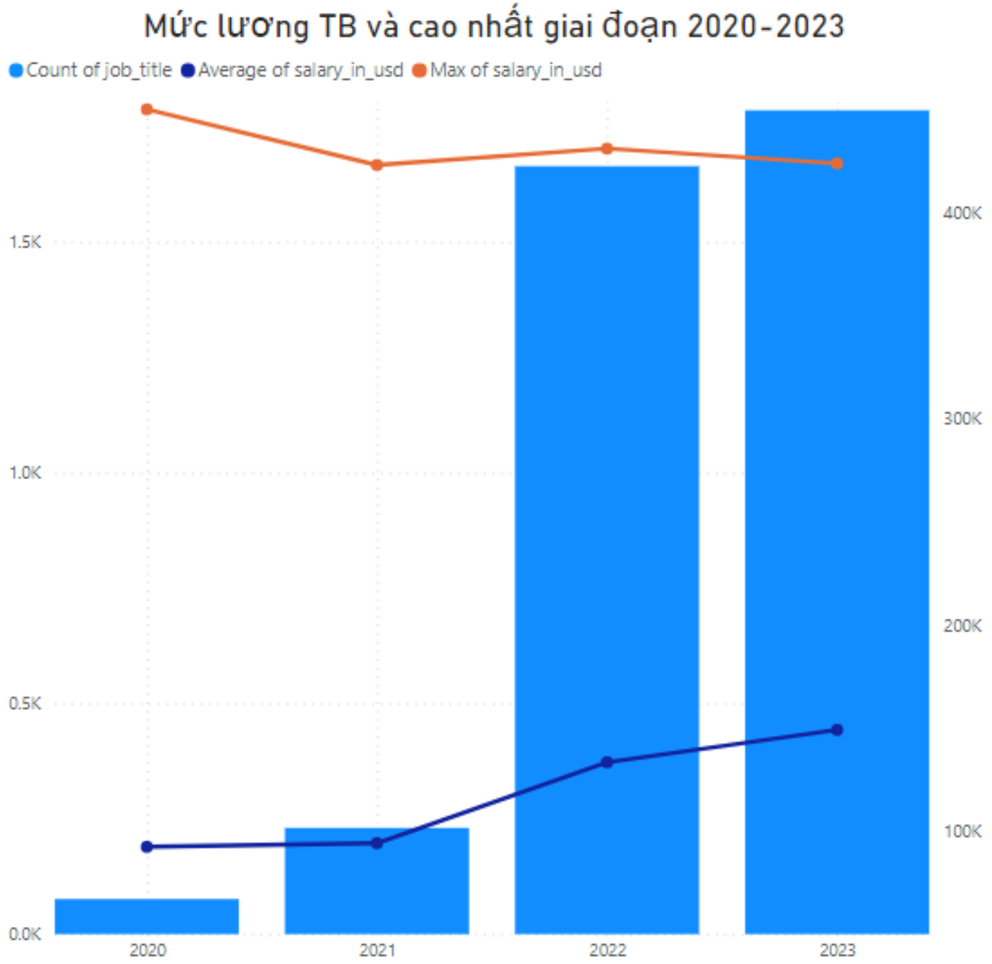
By Company size

Mid-sized companies account for the largest proportion of employees in the Data Science industry, which is the main growth driver for rapid job growth during the economic recovery period

Salary analysis

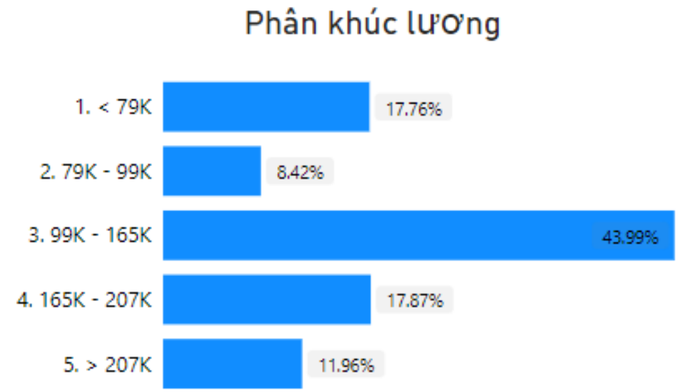
Data Science Industry
Period 2020 - 2023

By salary segment



23.62% **Average salary across the industry**

- Along with the number of jobs increasing from 76 (2020) to 1785 (2023), the average salary also recorded an increase of 23.62%. In particular, there will be a sharp increase in the period 2021-2022.
- However, the average salary and the highest salary in each year have a large difference, showing that the majority of jobs are recruiting offers in the average and low salary segment.
- In particular, the 99K - 165K segment accounts for the largest number of jobs.

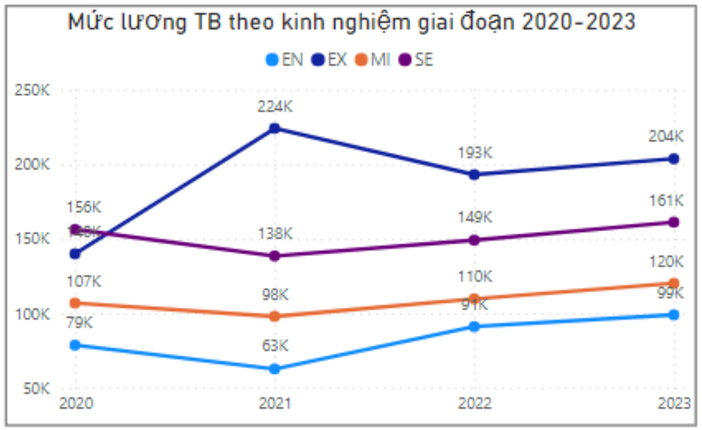


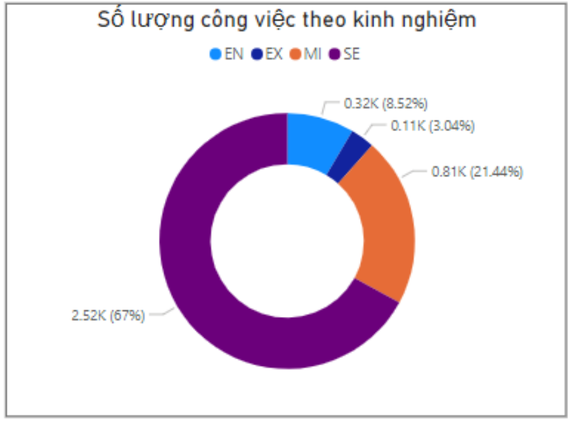
114.1K
2023

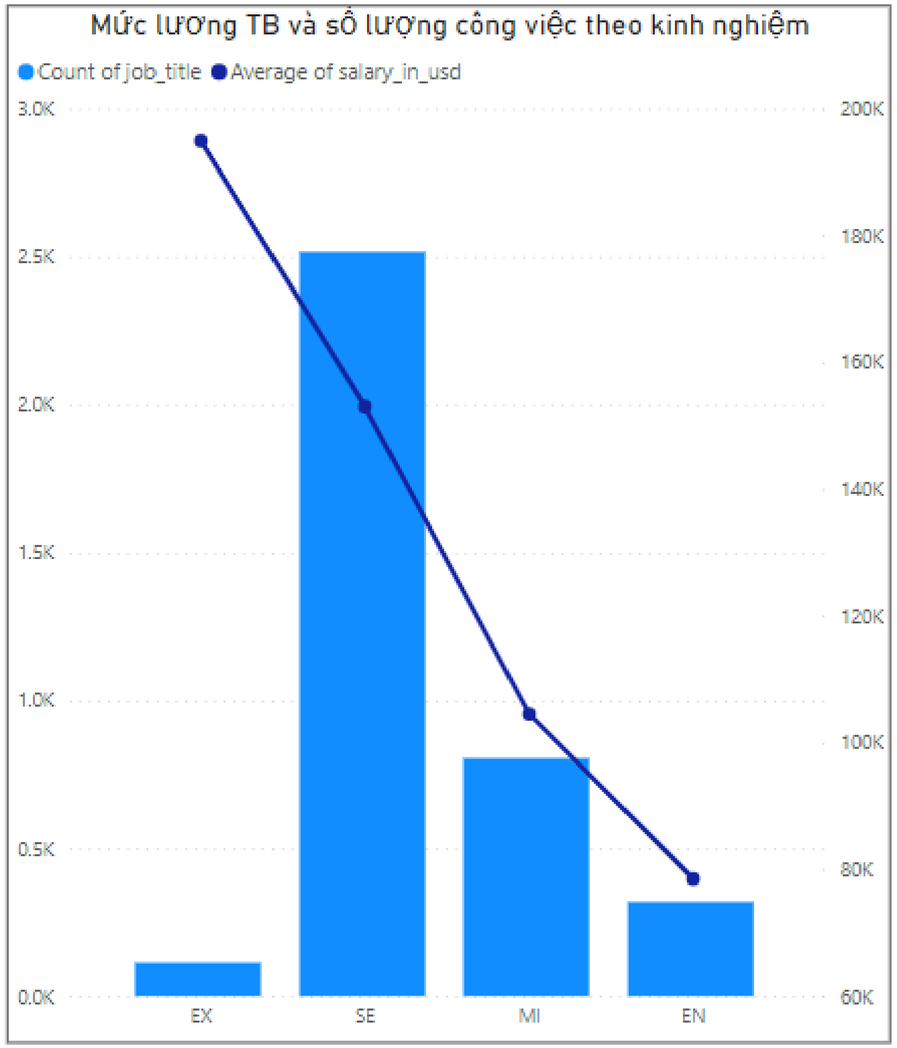
92.3K
2020

By Experience level

- Entry (EN): The number of EN jobs accounts for 8.52% in the period 2020-2023, with the average increase of 25.32%. As the lowest level of experience, EN's average salary has a large difference compared to the remaining experience levels.
- Mid-level (MI): The number of jobs ranks second in the entire market, with an average salary about 20% higher than EN. Shows that after accumulating 1-2 years of experience, workers can significantly increase their income.
- Senior (SE): Is the experience level with the highest recruitment demand (accounting for 67% of the market). However, the average salary does not have major fluctuations, and will even record a decrease in 2021.
- Expert (EX): The highest level of experience, with a small number of jobs and an impressively high average salary of 204K/year in 2023, up to nearly 400K/year.

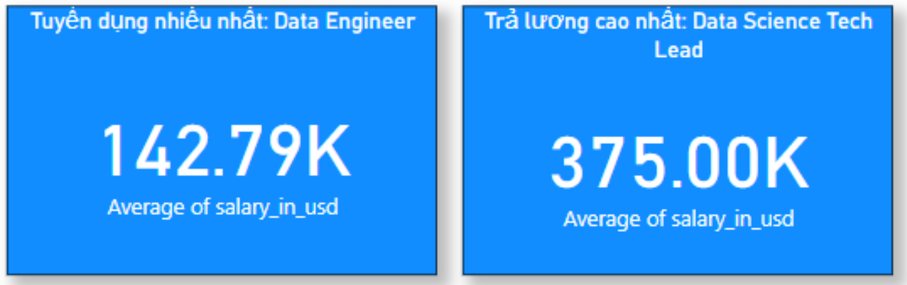


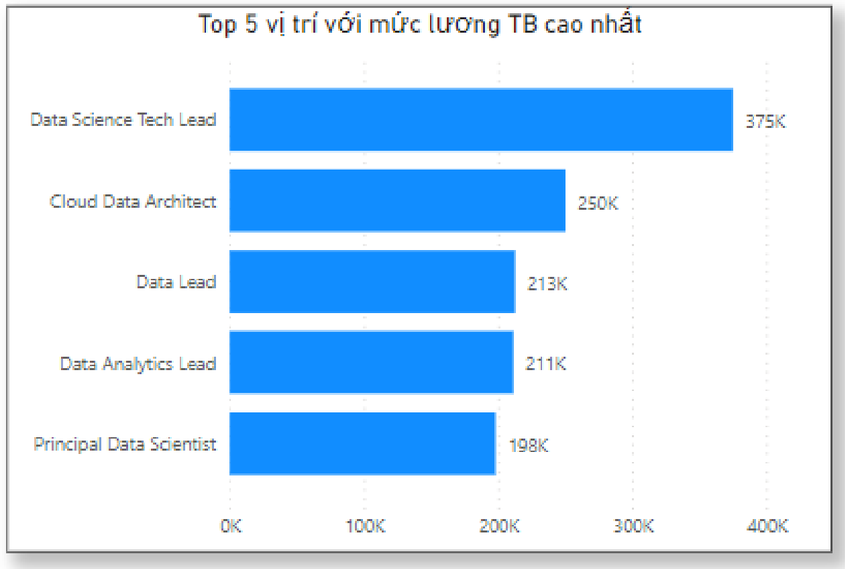


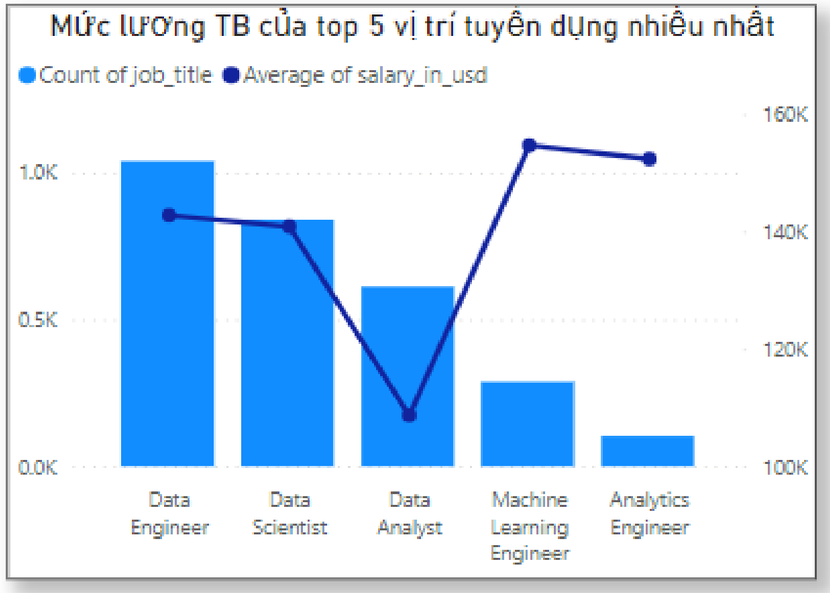


By Position

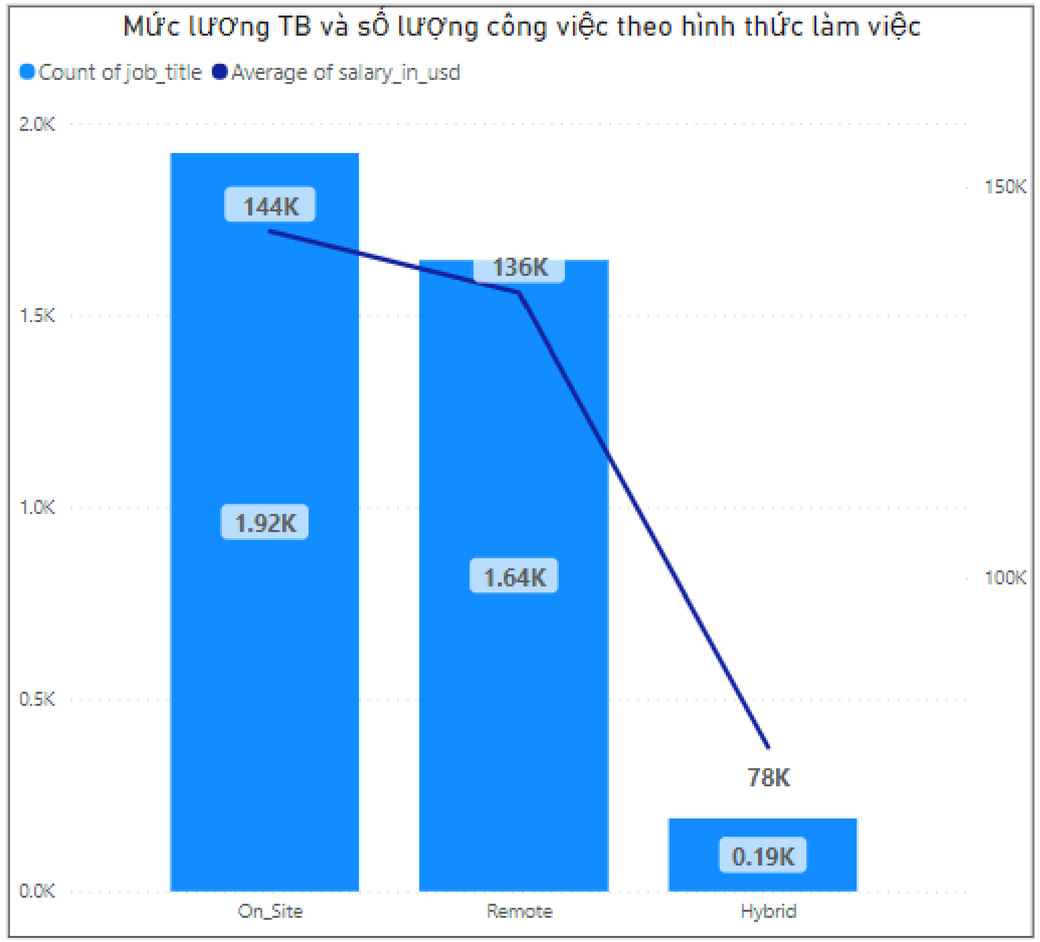
- The data set records 93 different job positions, of which the Data Engineer position is the most sought after by employers (142.79K USD/year) and the highest paid position is Data Science Tech Lead (375K USD/year).
- The position group with the highest average salary is in Lead positions. This can be easily explained because in addition to expertise, the Lead position requires the ability to manage human resources and a high level of responsibility.
- The most recruited position group is in diverse professional positions such as Data Engineer, Data Scientist, Data Analyst, ML Engineer,... Among them, the engineer group has the highest average salary.





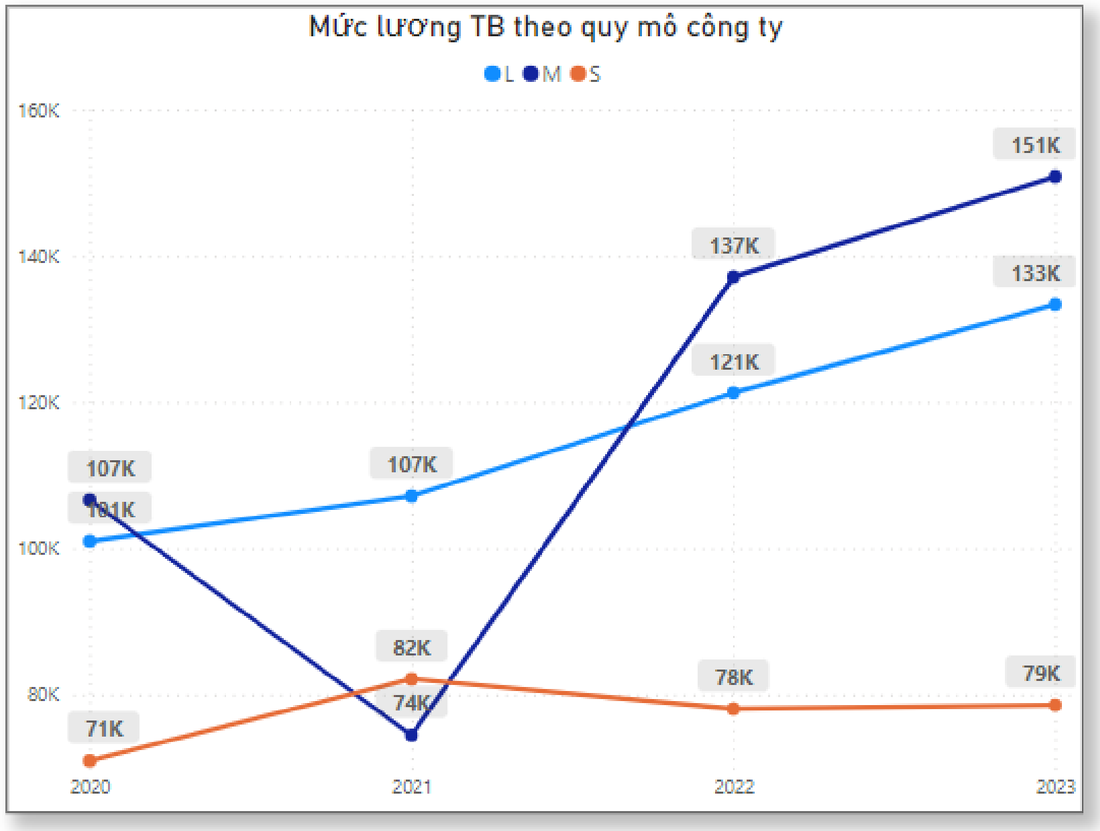


By Working type



- Remote: the number of remote jobs ranks 2nd by type of work. The average salary does not have much difference compared to the on-site form, showing that the market does not differentiate between direct and remote forms of work.
- On-site: Dominates the number of jobs and has the highest average salary. However, the year-over-year increase in average salaries shows signs of being slower than remote work, showing that the choice to work remotely tends to become a privilege of groups with higher average salaries.
- Hybird: 50/50 working form is the least popular form, average salary at 78K and number of jobs is 189 jobs in the period 2020-2023.

By Company size



- Small-scale companies (S): average salary for vacancies does not have large fluctuations, with 79K/year in 2023, showing that vacancies are often entry positions, with no department in charge yet formed. private data at small-scale companies.
- Medium-sized company (M): average salary in 2021 witnessed a decrease of about 30%, showing the impact of Covid-19 and the reduction in recruitment needs in the tourism industry. In the period 2022-2023, the salary will grow by nearly 100%, even exceeding the salary group of large-scale companies.
- Large-scale company (L): although the number of jobs will decrease by 50% in 2023, the average salary will still grow steadily, about 10.56%/year.

CONCLUSION

By salary segment

The 99K - 165K segment is the most popular average salary. Average salary increases sharply by 41% in 2022

By Experience level

Most employers recruit at the Senior experience level. The average salary of Mid-level and Entry groups does not have a big difference

By Working type

On-site form has the highest average salary, however the average salary for remote work is also close to the highest

By Company size

Medium-sized companies pay the highest average salary in the industry but are not stable. Large-scale company, operating a professional data system, maintaining a steady average salary increase of 10.56%/year

Labor market forecast

Data Science Industry In 2024

Labor market forecast

- Companies will continue to recruit more experienced personnel (from Mid to Senior Level) to save time and personnel training costs and receive higher work performance. This is a necessary factor to maintain business efficiency in the current economy and prepare for the next growth period.

- Large-scale companies will continue to provide jobs and opportunities to increase stable income for the workforce. The medium-sized company sector is more sensitive to economic trends and can maintain growth momentum, but will slow down compared to the previous period.

- Remote work will gradually become a concentrated benefit for highly specialized and high-income groups when the number of Remote jobs decreases sharply but the average income still increases. In the short term, On-site jobs still dominate because they are so deeply embedded in corporate culture.

Recommendations for employees

- Workers need to continuously update their knowledge and work skills. For highly specialized personnel, in order to reach the next step at the management level, it is necessary to add additional non-professional skills (e.g. management, leadership)
- Although the number of jobs at large-scale companies is not much, leading to high competition, workers should consider working here to secure the best career before economic fluctuations. For jobs at medium-sized companies, workers need to regularly monitor market fluctuations and the economy in general to be proactive with their own careers.
- Employees, if they put their career first, should consider working in large countries with strong demand for Data Science personnel.

Thank you for listening!