CoKaCoLa

Information available in audio.

Phân tích ngành Data ở thị trường Mỹ

2020 - 2023

Thành Công
Anh Khoa
Lan Anh
Biển Ngọc

CoKaCoLa

Nội dung chính

1. Lý do chọn đề tài
   - So sánh
   - Xác định mục tiêu
2. Làm sạch dữ liệu
   - Xử lý dữ liệu Null
   - Chia dữ liệu thành các bảng nhỏ
3. Trực quan hóa dữ liệu → đưa ra nhận xét từ các con số
4. Kết luận

Data processing

National Occupational Employment and Wage Estimates United States – May 2022

Data Science Salaries 2023

PYTHON VÀ SQL

```python
import pandas as pd

# Pandas settings for visualization
pd.set_option('display.max_columns', None)
pd.set_option('display.width', None)
pd.set_option('display.max_colwidth', None)

# Import csv using pandas
clean = pd.read_csv('national_M2022_dl.csv')
df1 = pd.DataFrame(clean)

# Cleaning data
# Drop unnecessary columns
df1.drop(columns=['AREA', 'AREA_TITLE', 'AREA_TYPE', 'PRIM_STATE',
                  'NAICS', 'NAICS_TITLE','I_GROUP', 'OWN_CODE',
                  'JOBS_1000','LOC_QUOTIENT', 'PCT_TOTAL', 'PCT_RPT',
                  'H_PCT10', 'H_PCT25', 'H_MEDIAN', 'H_PCT75',
                  'H_PCT90', 'A_PCT10','A_PCT25', 'A_MEDIAN',
                  'A_PCT75', 'A_PCT90', 'ANNUAL', 'HOURLY'],
        inplace=True)
```

Xử lý

```sql
WITH raw_table as(
    Select *
    ,cast(replace(replace([H_MEAN],'*','0'), ',', '') as float) as HMean
    ,cast(replace(replace([A_MEAN],'*','0'), ',', '') as float) as AMean
    ,cast(REPLACE([TOT_EMP],',','')as float) as TotalEmployment
    ,cast([MEAN_PRSE] as float) Meanprse
    FROM [dbo].[Salary1]
    )
,fact_table as (SELECT
[OCC_CODE]
,[OCC_TITLE]
,[O_GROUP]
,TotalEmployment
,[EMP_PRSE]
,Meanprse
,round(IIF(HMean=0,AMean/260/8,HMean),2) as HMean
,round(IIF(AMean=0,HMean*260*8,AMean),0) as AMean
,ROW_NUMBER() OVER (PARTITION BY left(OCC_CODE,2)
ORDER BY left(OCC_CODE,2)) AS NumRows
FROM raw_table
WHERE O_GROUP IN ('Major','Minor'))

SELECT OCC_CODE, OCC_TITLE, O_GROUP, TotalEmployment
, EMP_PRSE, Meanprse, Hmean,Amean
From fact_table
WHERE NumRows <=4
```

Kiểm tra

National Occupation - Phân tích

OCC_TITLE	O_GROUP	TOT_EMP	EMP_PRSE	JOBS_100(	LOC_QUO	PCT_TOTA	PCT_RPT	H_MEAN	A_MEAN
All Occupations	total	147,886,000	0					29.76	61,900
Management Occu	major	9,860,740	0.6					63.08	131,200
Top Executives	minor	3,618,820	0.3					62.04	129,050
Chief Executives	broad	199,240	0.9					118.48	246,440
Chief Executives	detailed	199,240	0.9					118.48	246,440
General and Opera	broad	3,376,680	0.3					59.07	122,860
General and Opera	detailed	3,376,680	0.3					59.07	122,860
Legislators	broad	42,890	2.1					*	71,100
Legislators	detailed	42,890	2.1					*	71,100
Advertising, Marke	minor	977,490	2					73.23	152,320
Advertising and Pr	broad	22,010	3.1					70.7	147,050
Advertising and Pr	detailed	22,010	3.1					70.7	147,050
Marketing and Sale	broad	864,970	2.1					73.79	153,470
Marketing Manage	detailed	328,570	1.1					76.1	158,280
Sales Managers	detailed	536,390	2.8					72.37	150,530
Public Relations ar	broad	90,510	1.5					68.56	142,610
Public Relations M	detailed	64,280	1.4					72.13	150,030
Fundraising Manag	detailed	26,240	2.7					59.83	124,450
Operations Special	minor	2,322,010	1.2					71.25	148,190
Administrative Ser	broad	353,550	1.5					54.06	112,440
Administrative Ser	detailed	236,570	2.1					55.59	115,640
Facilities Manager	detailed	116,980	0.6					50.95	105,970
Computer and Infc	broad	533,220	1.2					83.49	173,670
Computer and Infc	detailed	533,220	1.2					83.49	173,670
Financial Manager	broad	740,780	1.2					79.83	166,050
Financial Manager	detailed	740,780	1.2					79.83	166,050
Industrial Producti	broad	211,710	0.6					58.13	120,900

Thừa, trống dữ liệu

Dữ liệu Null

Tên cột không rõ ràng

National Occupation - xử lý

Bỏ các cột không cần thiết, không thể phục hồi dữ liệu

```
df1.drop(columns=['AREA', 'AREA_TITLE', 'AREA_TYPE', 'PRIM_STATE', 'NAICS', 'NAICS_TITLE',
       'I_GROUP', 'OWN_CODE','JOBS_1000',
       'LOC_QUOTIENT', 'PCT_TOTAL', 'PCT_RPT',
       'H_PCT10', 'H_PCT25', 'H_MEDIAN', 'H_PCT75', 'H_PCT90', 'A_PCT10',
       'A_PCT25', 'A_MEDIAN', 'A_PCT75', 'A_PCT90', 'ANNUAL', 'HOURLY'],inplace=True)
```

Làm rõ tên cột

```
df1.rename(columns = {'OCC_CODE':'Occupation Code', 'OCC_TITLE':'Occupation Title',
                      'O_GROUP':'Occupation Group', 'TOT_EMP':'Total Employee',
                      'EMP_PRSE':'Employee PRSE','H_MEAN':'Hour Mean', 'A_MEAN':'Annual Mean',
                      'MEAN_PRSE':'Mean PRSE'}, inplace = True)
```

Thêm dữ liệu vào các ô null (Hour Mean = Annual Mean/260/8)

```
i1 = df1.index
index = df1["Hour Mean"] == '*'
result1 = i1[index]
result1.tolist()
df1['Annual Mean']=df1['Annual Mean'].str.replace(',','')
df1.loc[result1, 'Hour Mean'] = df1.loc[result1,'Annual Mean'].astype('float')/260/8
df1['Hour Mean'] = df1['Hour Mean'].astype('float').map("{:,.2f}".format)
```

National Occupation - Kết quả

Occupational code	Occupation Title	Occupation group	Total Employees	Employee PRSE	Hour Mean	Annual Mean	Mean PRSE
00-0000	All Occupations	total	147,886,000	0.0	29.76	61,900	0.2
11-0000	Management Occupations	major	9,860,740	0.6	63.08	131,200	0.6
11-1000	Top Executives	minor	3,618,820	0.3	62.04	129,050	1.1
11-1010	Chief Executives	broad	199,240	0.9	118.48	246,440	1.6
11-1011	Chief Executives	detailed	199,240	0.9	118.48	246,440	1.6
11-1020	General and Operations Managers	broad	3,376,680	0.3	59.07	122,860	1.1
11-1021	General and Operations Managers	detailed	3,376,680	0.3	59.07	122,860	1.1
11-1030	Legislators	broad	42,890	2.1	34.18	71,100	1.7
11-1031	Legislators	detailed	42,890	2.1	34.18	71,100	1.7
11-2000	Advertising, Marketing, Promotions, Public Relations, and Sales Managers	minor	977,490	2.0	73.23	152,320	0.4
11-2010	Advertising and Promotions Managers	broad	22,010	3.1	70.70	147,050	2.2
11-2011	Advertising and Promotions Managers	detailed	22,010	3.1	70.70	147,050	2.2
11-2020	Marketing and Sales Managers	broad	864,970	2.1	73.79	153,470	0.4
11-2021	Marketing Managers	detailed	328,570	1.1	76.10	158,280	0.6
11-2022	Sales Managers	detailed	536,390	2.8	72.37	150,530	0.4
11-2030	Public Relations and Fundraising Managers	broad	90,510	1.5	68.56	142,610	0.7
11-2032	Public Relations Managers	detailed	64,280	1.4	72.13	150,030	0.8
11-2033	Fundraising Managers	detailed	26,240	2.7	59.83	124,450	1.3
11-3000	Operations Specialties Managers	minor	2,322,010	1.2	71.25	148,190	0.4
11-3010	Administrative Services and Facilities Managers	broad	353,550	1.5	54.06	112,440	0.3
11-3012	Administrative Services Managers	detailed	236,570	2.1	55.59	115,640	0.4
11-3013	Facilities Managers	detailed	116,980	0.6	50.95	105,970	0.4
11-3020	Computer and Information Systems Managers	broad	533,220	1.2	83.49	173,670	0.4

Note: PRSE = Percentage relative standard error (% Sai số)

Occupation Title	Occupation Group	Hour Mean	Annual Mean	Mean PRSE
Management Occupations	major	63.08	131,200	0.6
Advertising, Marketing, Promotions, Public Relations, and Sales Managers	minor	73.23	152,320	0.4
Operations Specialties Managers	minor	71.25	148,190	0.4
Top Executives	minor	62.04	129,050	1.1
Business and Financial Operations Occupations	major	41.39	86,080	0.3
Financial Specialists	minor	44.37	92,290	0.5
Business Operations Specialists	minor	40.04	83,280	0.2
Computer and Mathematical Occupations	major	51.99	108,130	0.5
Mathematical Science Occupations	minor	52.22	108,620	0.5
Computer Occupations	minor	51.97	108,100	0.5
Architecture and Engineering Occupations	major	45.52	94,670	0.5

National Occupation

Hướng đi

Occupation Title	Occupation Group	Total Empl	Employee PRSE
Management Occupations	major	9,860,740	0.6
Top Executives	minor	3,618,820	0.3
Other Management Occupations	minor	2,942,420	0.6
Operations Specialties Managers	minor	2,322,010	1.2
Business and Financial Operations Occupations	major	9,677,720	0.7
Business Operations Specialists	minor	6,660,810	0.7
Financial Specialists	minor	3,016,910	0.7
Computer and Mathematical Occupations	major	5,003,910	0.2
Computer Occupations	minor	4,677,500	0.2
Mathematical Science Occupations	minor	326,400	0.9
Architecture and Engineering Occupations	major	2,481,170	0.3

Data Science Salaries – Phân tích

work_year	experien	employment_type	job_title	salary_in_usd	employee_residence	remote_ratio	company_location
2023	SE	FT	Principal Data Scien	85847	ES	100	ES
2023	MI	CT	ML Engineer	30000	US	100	US
2023	MI	CT	ML Engineer	25500	US	100	US
2023	SE	FT	Data Scientist	175000	CA	100	CA
2023	SE	FT	Data Scientist	120000	CA	100	CA
2023	SE	FT	Applied Scientist	222200	US	0	US
2023	SE	FT	Applied Scientist	136000	US	0	US
2023	SE	FT	Data Scientist	219000	CA	0	CA
2023	SE	FT	Data Scientist	141000	CA	0	CA
2023	SE	FT	Data Scientist	147100	US	0	US
2023	SE	FT	Data Scientist	90700	US	0	US
2023	SE	FT	Data Analyst	130000	US	100	US
2023	SE	FT	Data Analyst	100000	US	100	US
2023	EN	FT	Applied Scientist	213660	US	0	US
2022	SE	FT	Principal Machine L	190000	US	100	US
2022	SE	FT	Data Engineer	194000	US	100	US
2022	SE	FT	Data Engineer	129400	US	100	US
2022	SE	FT	Data Analyst	201000	US	100	US
2022	SE	FT	Data Analyst	89200	US	100	US
2022	SE	FT	Data Scientist	165000	US	0	US
2022	SE	FT	Data Scientist	125000	US	0	US
2022	SE	FT	Applied Scientist	230000	US	100	US
2022	SE	FT	Applied Scientist	196000	US	100	US
2022	MI	FT	Machine Learning E	130000	US	0	US
2022	MI	FT	Machine Learning E	90000	US	0	US
2022	MI	FT	Machine Learning R	150000	US	100	US

Nhiều kí hiệu không rõ ràng, trực quan

Không có khóa chính

Phân chia lẫn lộn theo năm, bị trùng

Data Science Salaries - xử lý

Tạo dictionary sau đó tiền hành thay vào các cột

```
Experience_dict={'SE':'Senior','MI':'Middle','EN':'Entry','EX':'Experience'}
Employment_dict={'FT':'Full Time', 'CT':'Casual Time', 'FL':'Freelancer', 'PT':'Part Time'}
National_dict ={'ES':'Spain', 'US':'United States of America', 'CA':'Canada',
                'DE':'Germany', 'GB':'United Kingdom of Great Britain and Northern Ireland',
                'NG':'Nigeria', 'IN':'India', 'HK':'Hong Kong', 'PT':'Portugal', 'NL':'Netherlands',
                'CH':'Switzerland', 'CF':'Central African Republic', 'FR':'France', 'AU':'Australia',
```

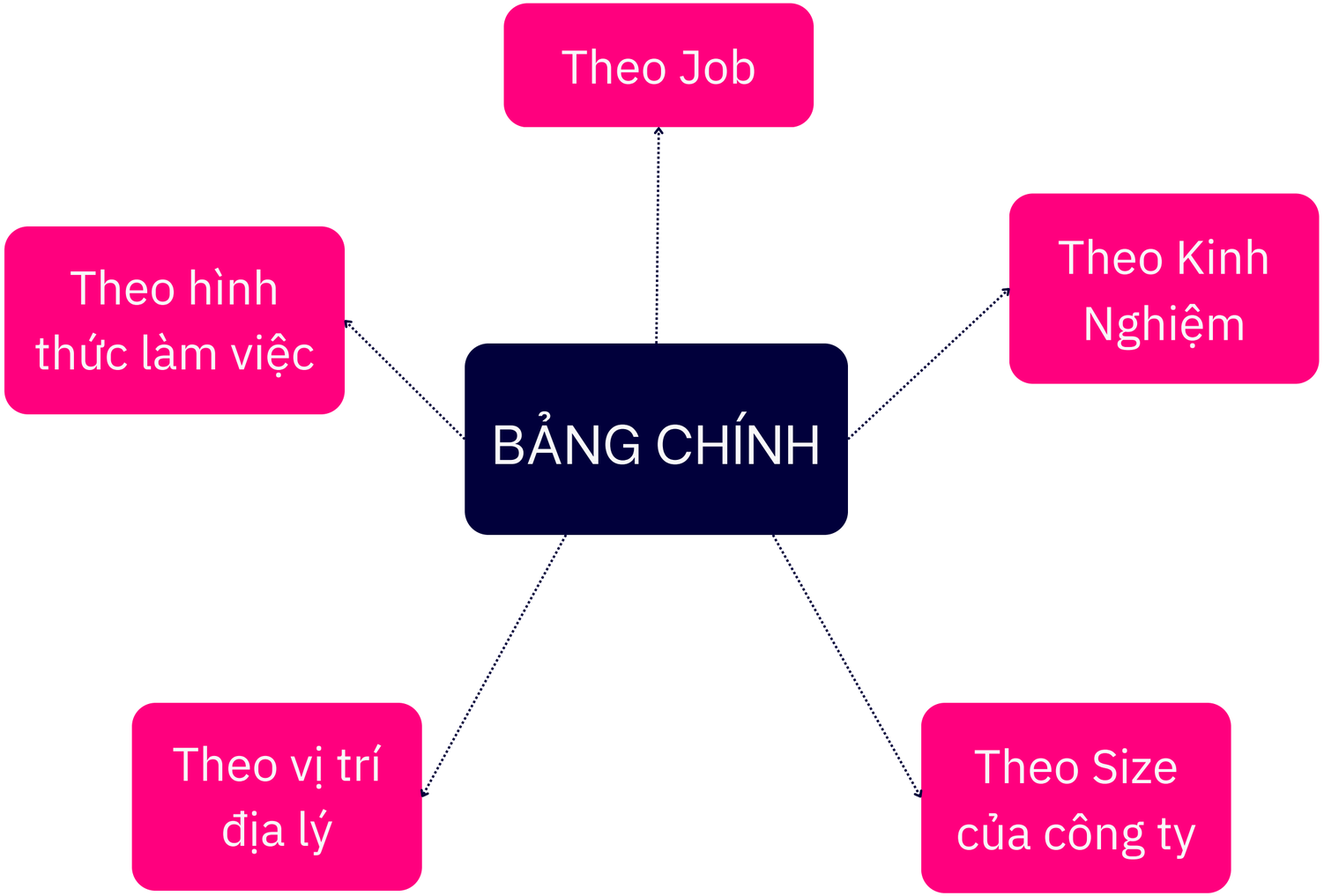
Dùng pandasql để lọc dữ liệu xong tiến hành pivot

```
By_Job_Title = ps.sqldf\
    ('select work_year,job_title, sum(salary_in_usd) '
     'from df group by work_year,job_title '
     'order by salary_in_usd',locals())

# Change format of work_year data from 2020.0 to 2020
By_Job_Title['work_year']=By_Job_Title['work_year'].astype(str).str.replace('.0','')
# Drop null rows
By_Job_Title = By_Job_Title.dropna()
# Pivot
By_Job_Title=By_Job_Title.pivot(index='job_title',columns='work_year',values='sum(salary_in_usd)')
# Add one more column Total
By_Job_Title['Total']= np.nansum((By_Job_Title['2020'],By_Job_Title['2021'],By_Job_Title['2022'],By_Job_Title['2023']
By_Job_Title.to_csv('By_Job_Title.csv',index = True)
```

Data Science Salaries

Hướng đi



Data Science Salaries- Kết quả

job_title	2020	2021	2022	2023	Total
3D Computer Vision Researcher		25409	60000		85409
AI Developer			581304	922023	1503327
AI Programmer			40000	70000	110000
AI Scientist					

company_location	2020	2021	2022	2023	Total
Algeria			100000		100000
American Samoa		38053	50000		88053
Argentina			75000		75000
Armenia				50000	50000

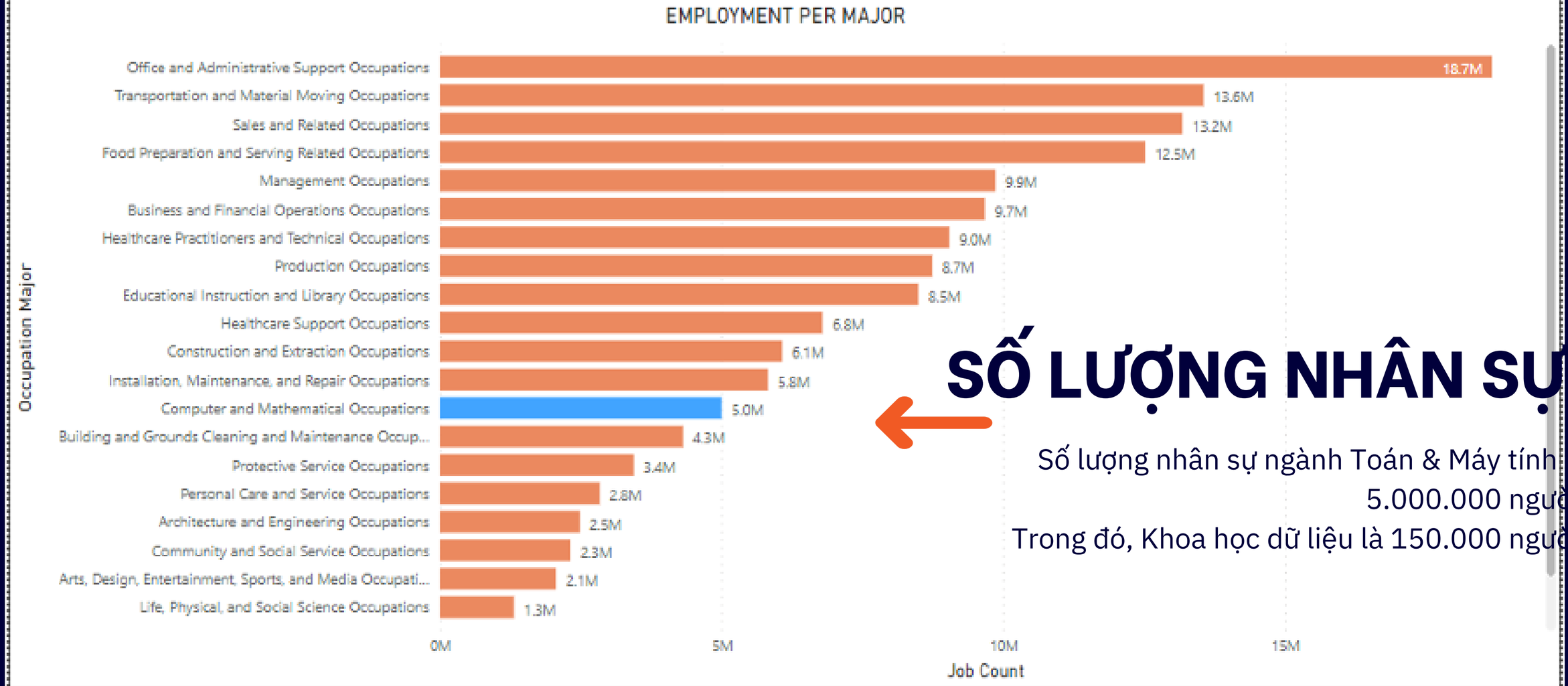
employment_type	2020	2021	2022	2023	Total
Casual Time/Contract	100000	791000	187969	55500	1134469
Freelancer	60000	77555	280523	100000	518078
Full Time	6817365	20614317	220947500	265873012	514252194
Part Time					

company_size	2020	2021	2022	2023	Total
Large	3535013	13610044	23892546	12671043	53708646
Middle	1706023	4169930	193768173	251646493	451290619
Small	1773964	3860084	4214746	1728755	11577549

experience_level	2020	2021	2022	2023	Total
Entry	1322767	3019789	9548747	11243508	25134811
Experienced	419833	1861280	7718672	12222341	22222126
Middle	2802071	7554758	36571321	37215231	84143381
Senior	2470329	9204231	168036725	205365211	385076496

Trực quan hoá bằng Power BI

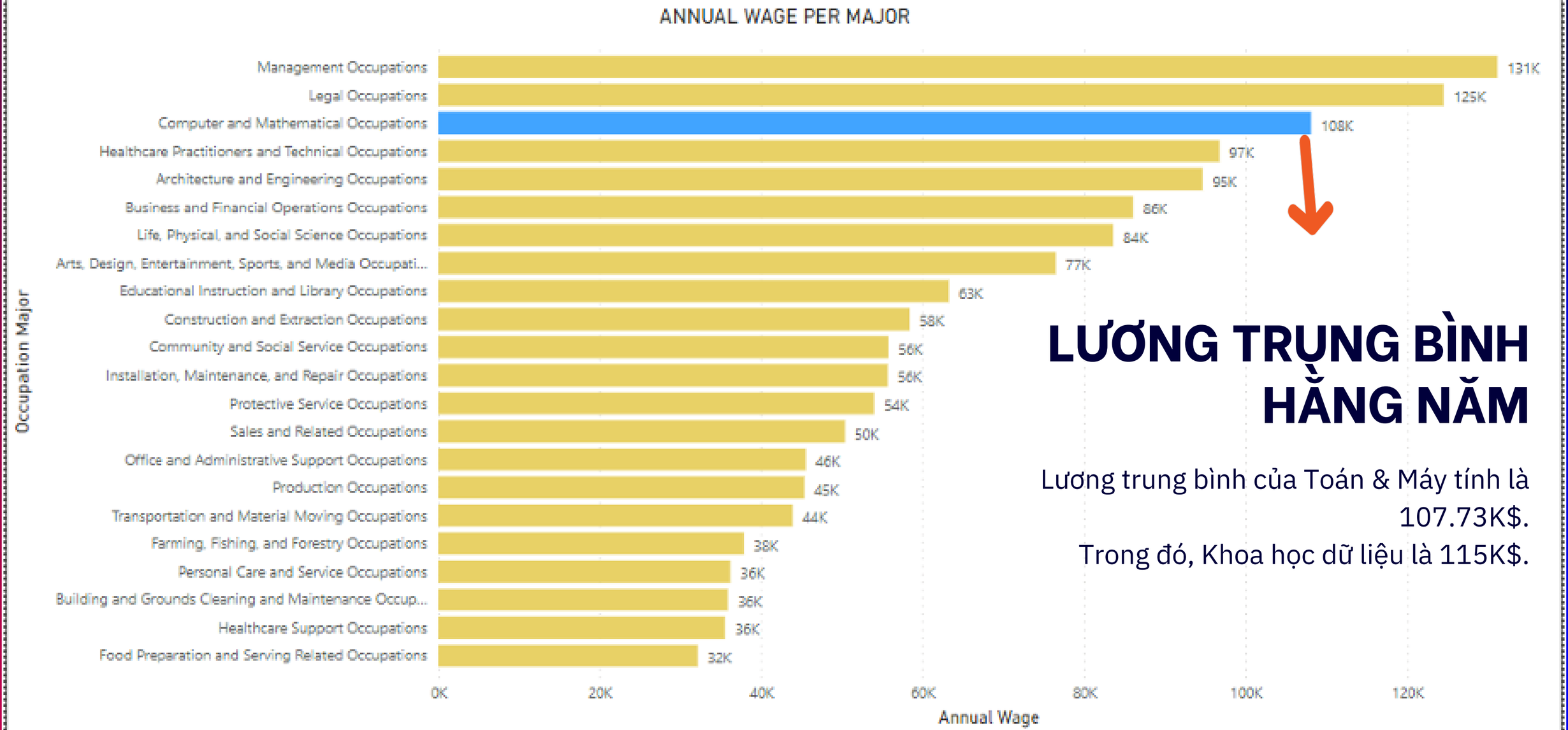
1	Tương quan Nhân sự và Mức lương ngành Data so với các ngành khác
2	Tổng quan các công việc ngành Data & Top các công việc phổ biến
3	Các yếu tố ảnh hưởng tới Số lượng nhân sự & Lương ngành Data
4	Tổng kết



SỐ LƯỢNG NHÂN SỰ

Số lượng nhân sự ngành Toán & Máy tính
5.000.000 ngườ
Trong đó, Khoa học dữ liệu là 150.000 ngườ

22	5M	160K
Total Occupation Major	Computer & Math Employment	Data Scientist Employment
148M	3.38%	0.11%
Total Employment	Computer & Math Employment Percentage	Data Scientist Employment Percentage



LƯƠNG TRUNG BÌNH HẰNG NĂM

Lương trung bình của Toán & Máy tính là 107.73K$.

Trong đó, Khoa học dữ liệu là 115K$.

65.98K

Average Annual Wage

107.73K

Computer & Math Occupation Average Annual Wage

115K

Data Scientist Average Annual Wage

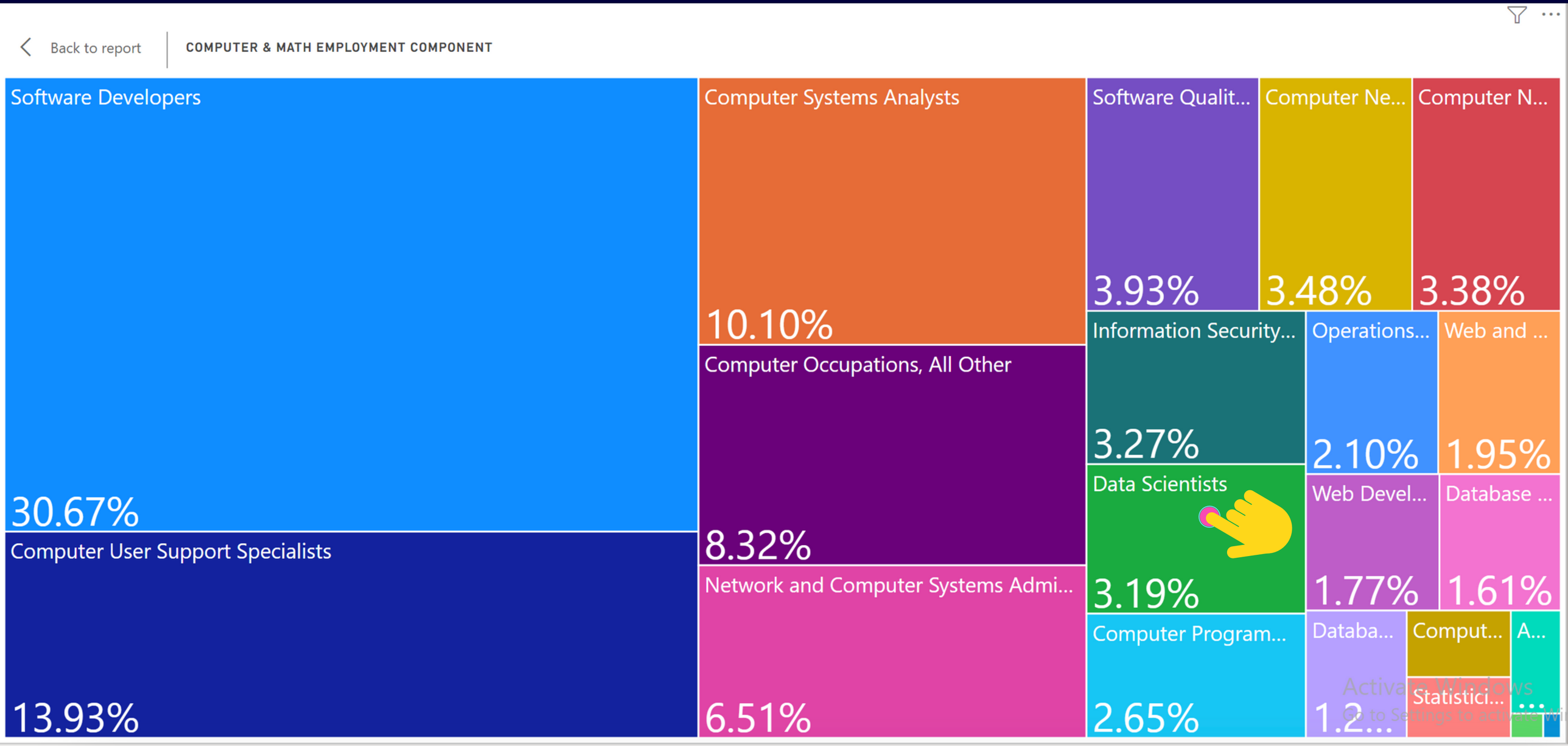
Trực quan hoá bằng Power BI

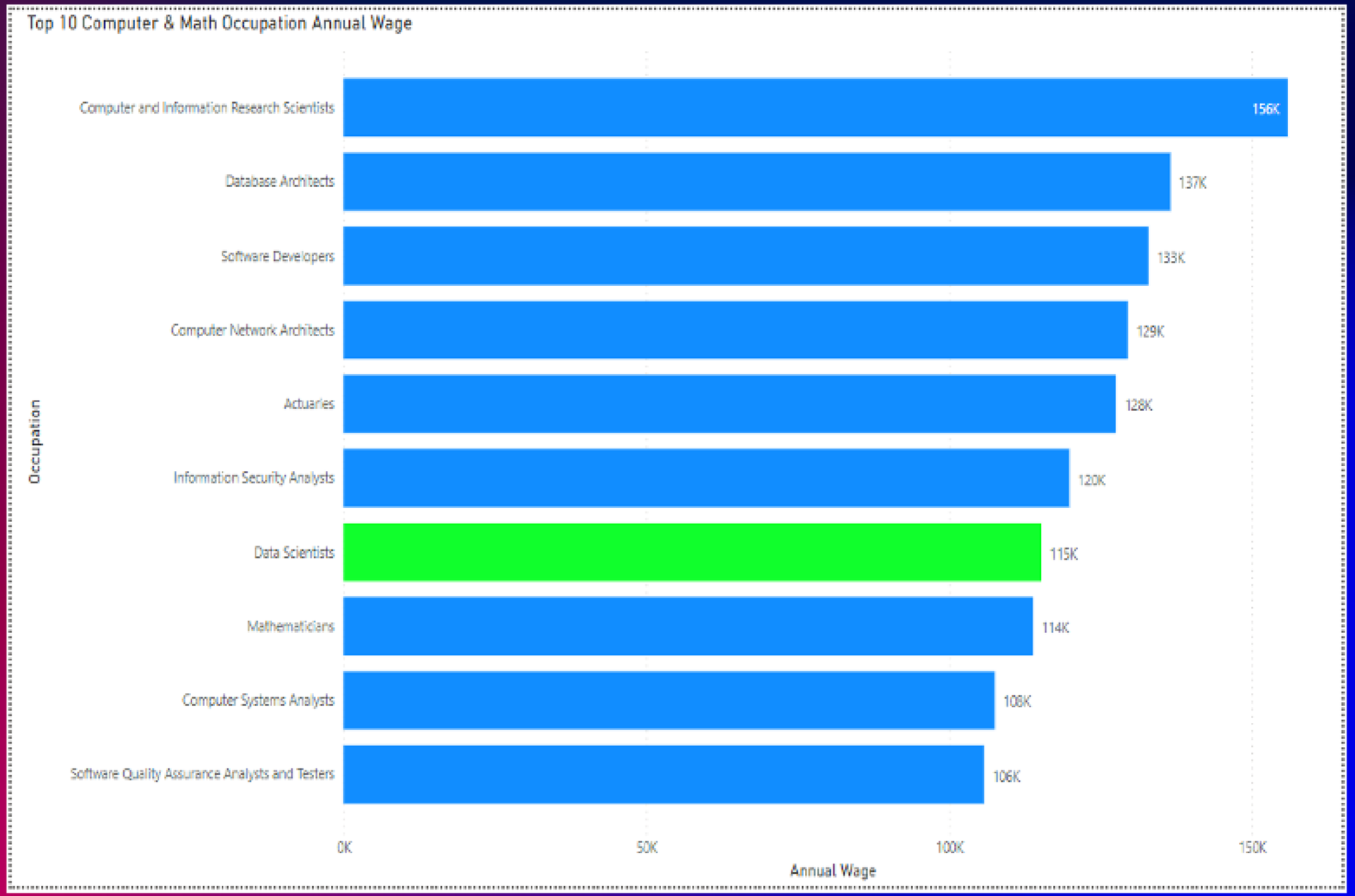
Số lượng nhân sự còn ít, nhu cầu tuyển dụng cao

Lương cao!!!

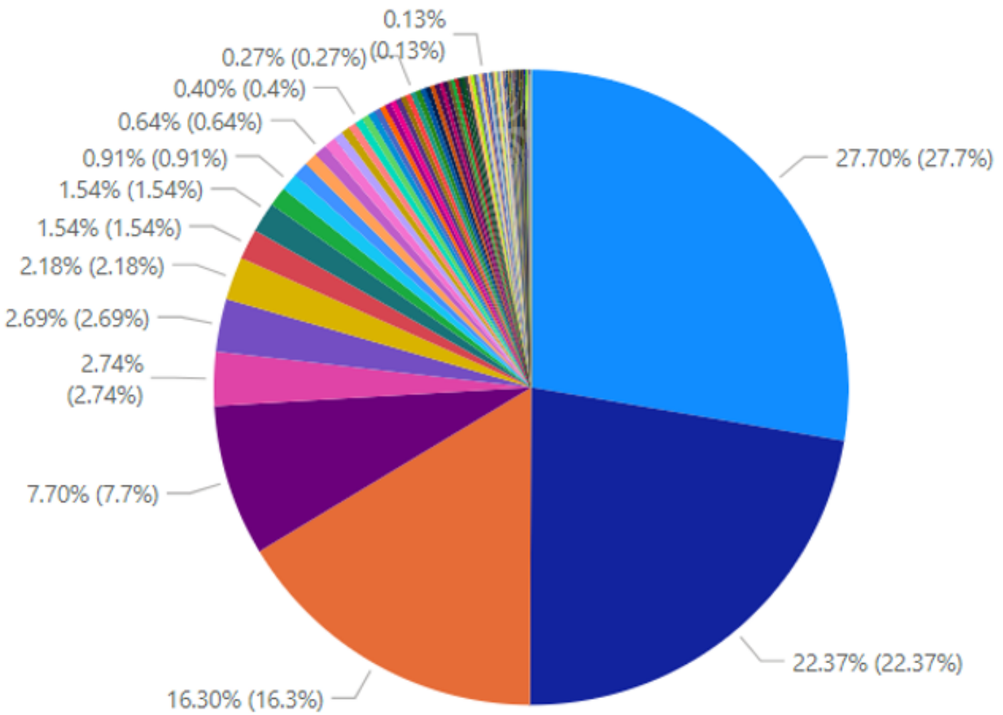
Ngành hấp dẫn

Trực quan hoá bằng Power BI




Top 10 Computer & Math Occupation Annual Wage
Computer and Information Research Scientists
156K
Database Architects
137K
Software Developers
133K
Computer Network Architects
129K
Actuaries
128K
Information Security Analysts
120K
Data Scientists
115K
Mathematicians
114K
Computer Systems Analysts
108K
Software Quality Assurance Analysts and Testers
106K
Occupation
0K
50K
100K
150K
Annual Wage

Data Job Popularity 2020-2023

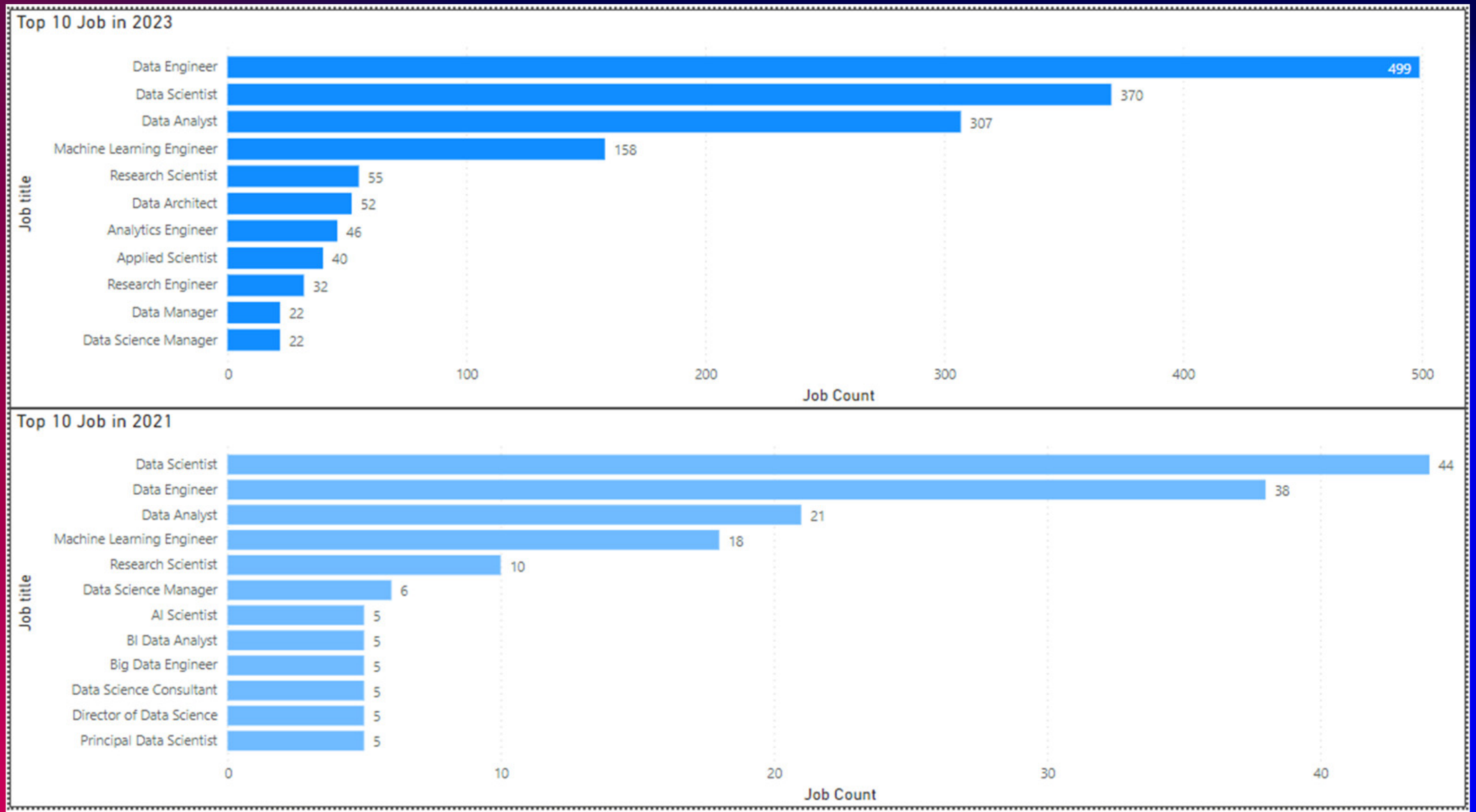


Job title

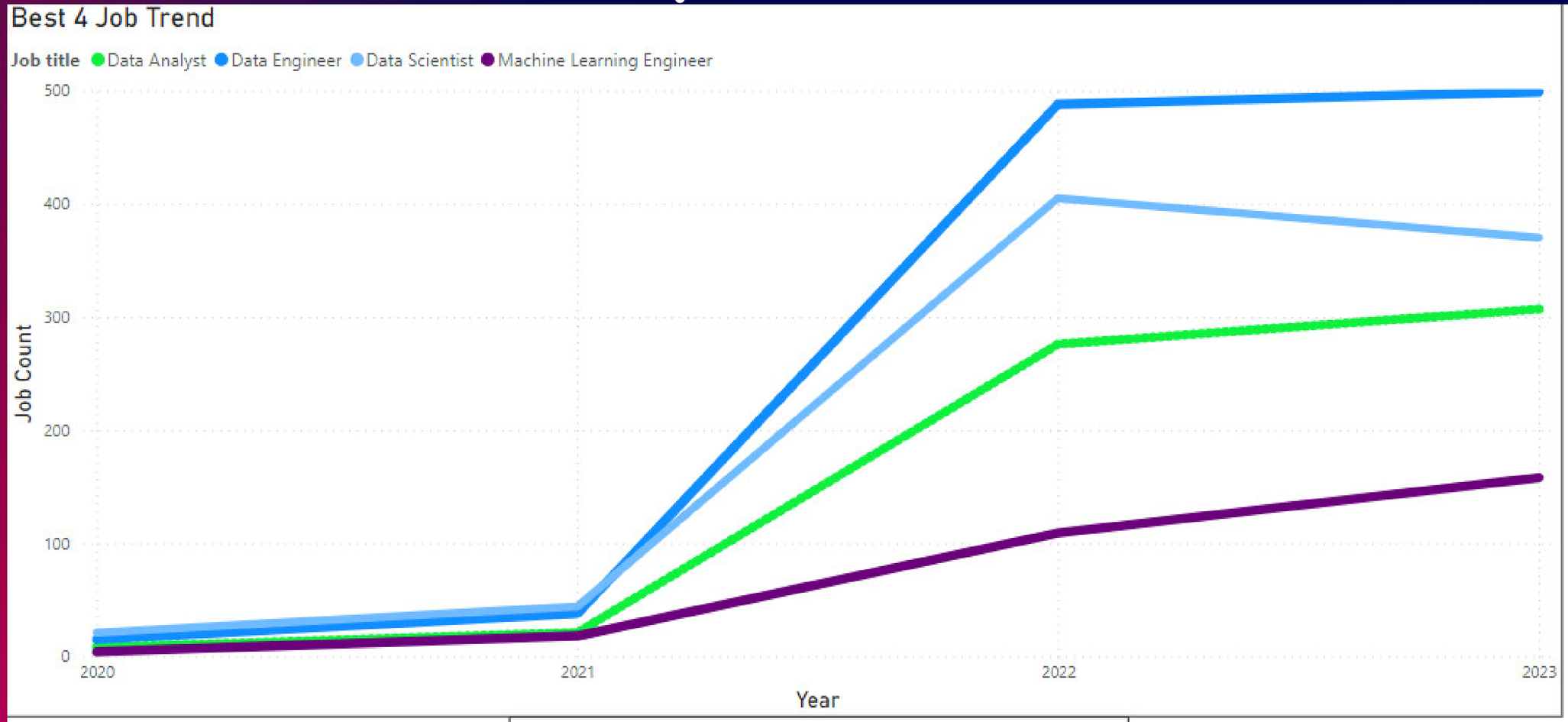
- Data Engineer
- Data Scientist
- Data Analyst
- Machine Learning Engineer
- Analytics Engineer
- Data Architect
- Research Scientist
- Applied Scientist
- Data Science Manager
- Research Engineer
- ML Engineer
- Data Manager
- Machine Learning Scientist
- Data Science Consultant
- Data Analytics Manager
- Computer Vision Engineer
- AI Scientist
- BI Data Analyst
- Business Data Analyst
- Data Specialist
- BI Developer
- Applied Machine Learning Scientist
- AI Developer
- Big Data Engineer
- Director of Data Science

Job title	Job Count
Data Engineer	1040
Data Scientist	840
Data Analyst	612
Machine Learning Engineer	289
Analytics Engineer	103
Data Architect	101
Research Scientist	82
Applied Scientist	58
Data Science Manager	58
Research Engineer	37
ML Engineer	34
Data Manager	29
Machine Learning Scientist	26
Data Science Consultant	24
Data Analytics Manager	22
Computer Vision Engineer	18
AI Scientist	16
BI Data Analyst	15
Business Data Analyst	15
Data Specialist	14
Total	**3433**

Check top 10 các Job Title phổ biến nhất năm 2021 (giai đoạn Covid) và 2023



Xu hướng Top 4 các Job phổ biến nhất trong suốt giai đoạn 2020-2023



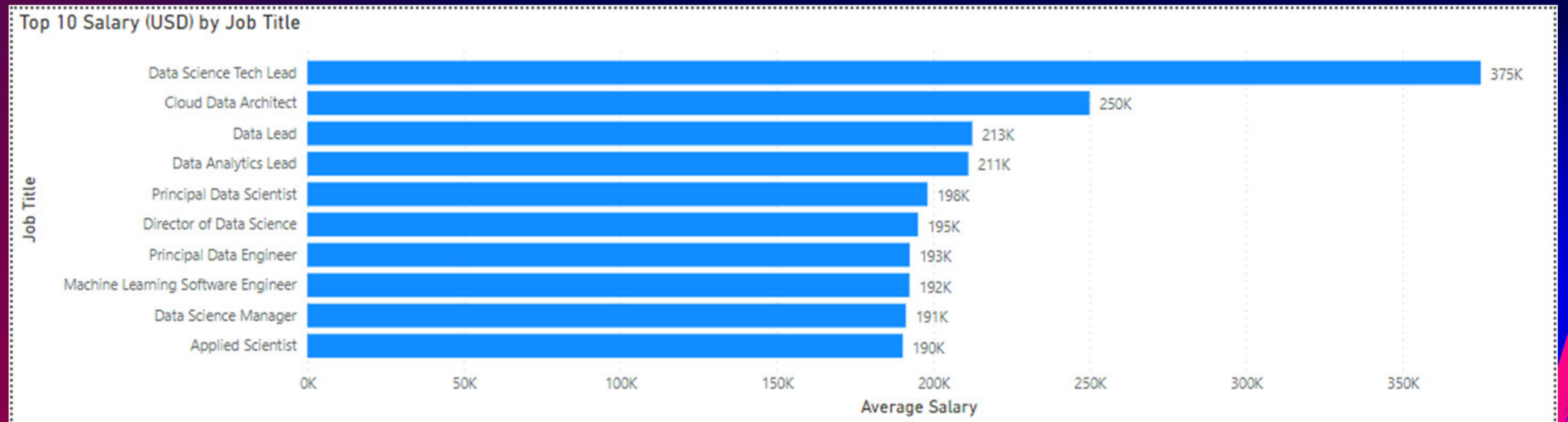
Job Title	2020	2021	2022	2023	Total
Data Engineer	15	38	488	499	**1040**
Data Scientist	21	44	405	370	**840**
Data Analyst	8	21	276	307	**612**
Machine Learning Engineer	4	18	109	158	**289**
Total	**48**	**121**	**1278**	**1334**	**2781**

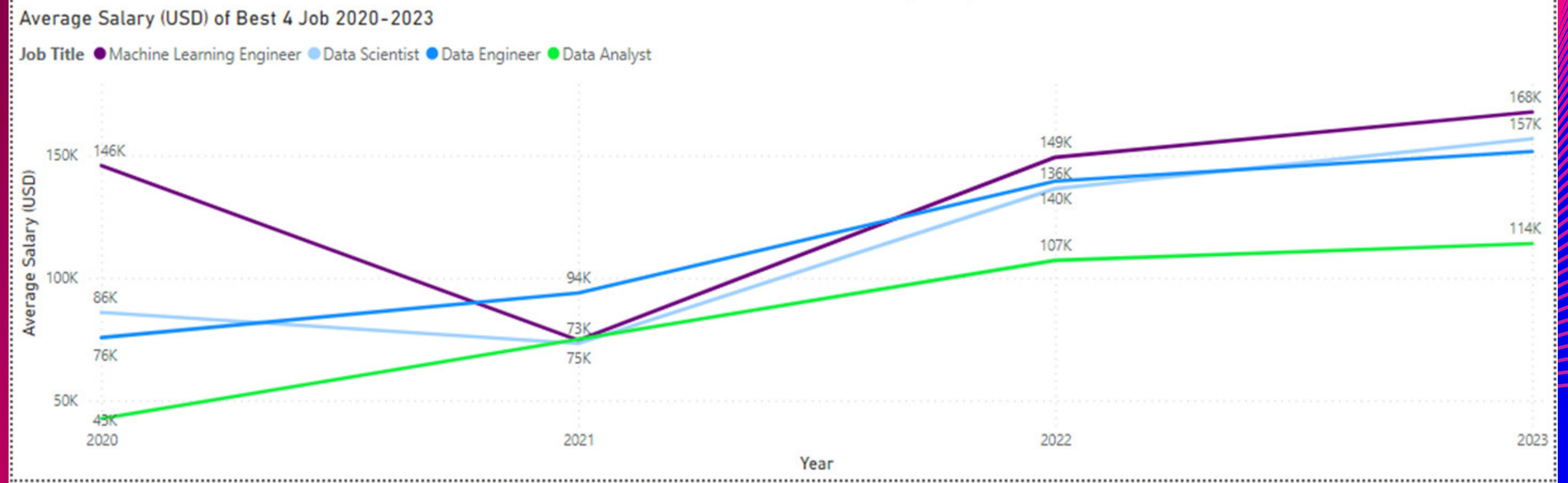
Trực quan hoá bằng Power BI

Số lượng nhân sự tăng lên đáng kể ở các Công việc phổ biến kể từ năm 2021, và vẫn còn tăng lên.

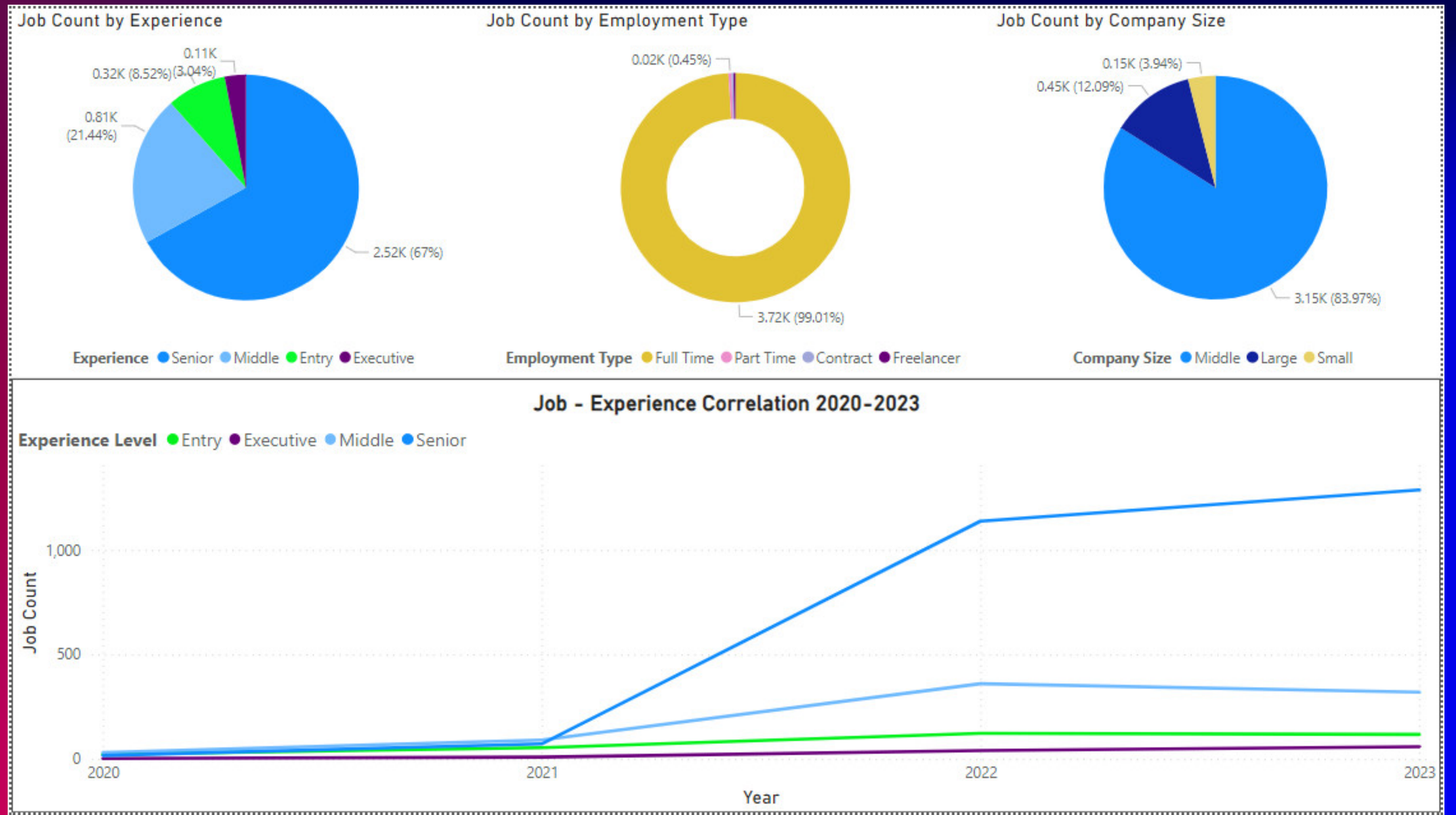
Quan trọng: Bảng số liệu chỉ tính đến tháng 4/2023.

Top các Công việc với lương USD cao nhất, và mức lương Top 4 phổ biến nhất 2020-2023

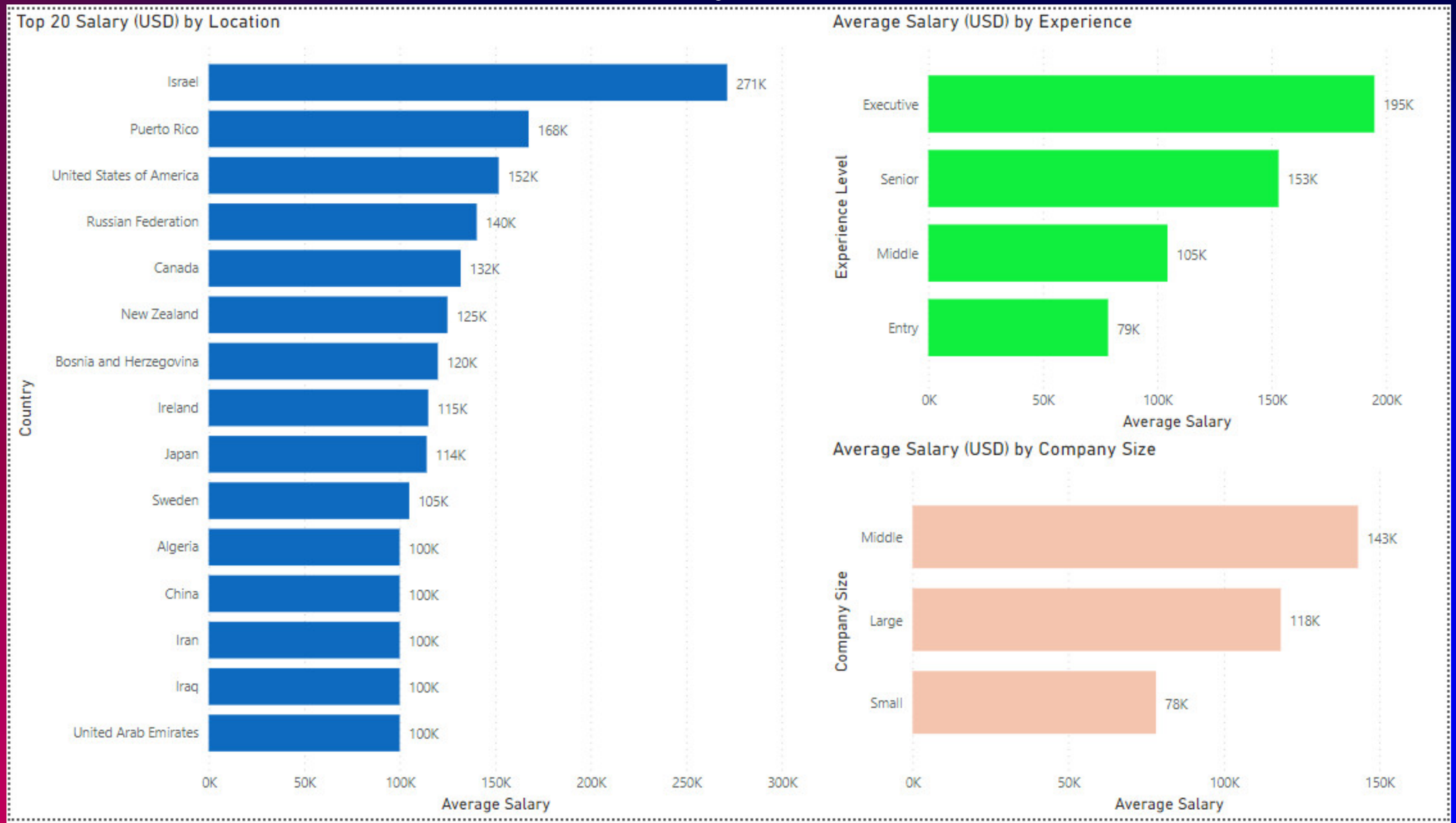




Tương quan Kinh nghiệm, Thời gian làm việc và Quy mô công ty với Số lượng công việc



Tương quan Mức lương với Địa điểm công ty, Kinh nghiệm và Quy mô



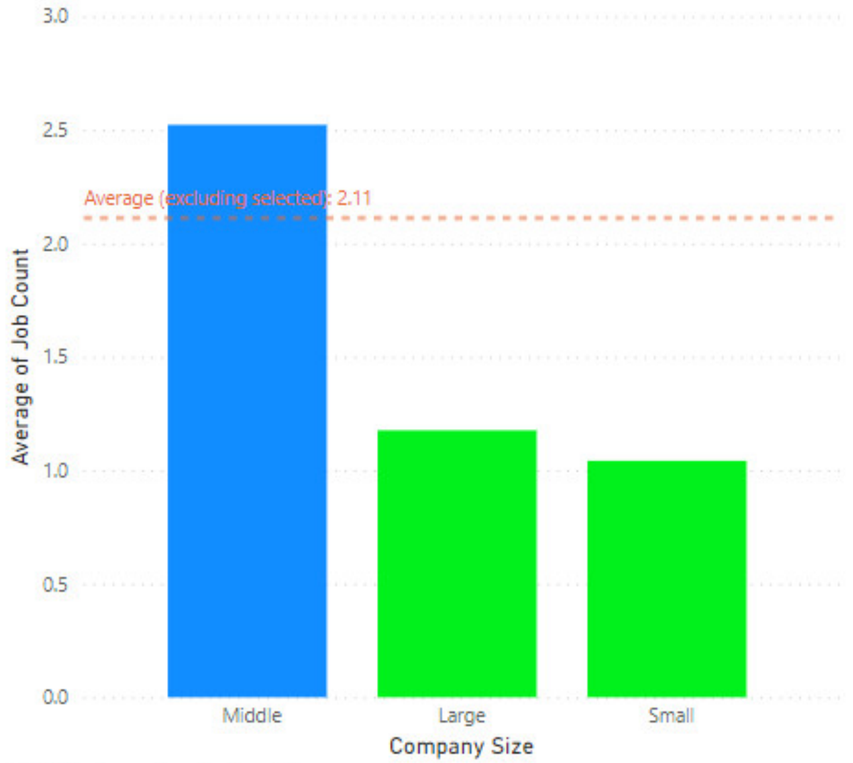
Key Factors to affect Job Count

Key influencers

What influences Job Count to Increase ?

When...	....the average of Job Count increases by
Company Size is Middle	1.38
Company Location is United States of America	1.36
Experience Level is Senior	1.12
Salary in USD is more than 64980	1.1

← Job Count is more likely to increase when Company Size is Middle than otherwise (on average).



☐ Only show values that are influencers

Analyze

Job Count

Explain by

Company Location

Company Size

Experience Level

Salary (USD)

Expand by

work_year

employment_type

KẾT LUẬN

Nhu cầu sử dụng dữ liệu trong công việc ngày càng cao:
+ Lượng dữ liệu thu thập được cũng ngày càng nhiều.
+ Doanh nghiệp cần dữ liệu để đưa ra chiến lược kinh doanh.
Nguồn lực còn khan hiếm => Cơ hội phát triển cao ở nhiều lĩnh vực, trên nhiều quốc gia.

NGƯỜI LÀM DATA