

# Báo cáo đồ án thực hành cuối kì

Môn học: Lập trình cho Khoa học dữ liệu

Giáo viên hướng dẫn:

1. Lê Nhật Nam
2. Trần Đại Chí

# Nhóm 13

Họ và tên

MSSV

**Trần Thông Lực**

**20120530**

**Phùng Hữu Tài**

**20120571**

**Lê Quang Thọ**

**20120588**

**Nguyễn Anh Tuấn**

**20120614**

# A. Thu thập dữ liệu

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

### 1. Chủ đề

Data Scientist Jobs

### 2. Lý do lựa chọn chủ đề

- Nhóm muốn tìm hiểu kỹ hơn về Khoa học dữ liệu: về nhu cầu và vị trí tuyển dụng, các công ty phù hợp, lĩnh vực, lương, ...

### 3. Dữ liệu

- Nguồn dữ liệu: <https://www.kaggle.com/datasets/andrewmvd/data-scientist-jobs>
- Được tạo bởi: <https://github.com/picklesueat/>
- Phương pháp thu thập dữ liệu:
  - Cào dữ liệu từ trang web tuyển dụng: <https://www.glassdoor.com/Job/> với từ khóa DataScientist
  - Giấy phép: được cấp phép theo MIT License

# A.Thu thập dữ liệu

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

### 1. Thời điểm thu thập dữ liệu

- 23/11/2022
- 3909 dòng, 17 cột

### 2. Cấu trúc tập dữ liệu

Column	Meaning	Column	Meaning	Column	Meaning
Unnamed: 0	ID	Company name	Company name	Industry	Industry
index	same as ID	Location	Job location	Sector	Sector within industry
Job Title	Job Title	Headquarters	Company headquarters location	Revenue	Yearly revenue
Salary Estimate	Salary Estimate	Size	Company size	Competitors	Main competitor
Job Description	Job Description	Founded	Year company was founded	Easy Apply	Whether easy apply is available or not
Rating	Company rating	Type of ownership	Type of company ownership		

# B. Khám phá dữ liệu



## I. Câu hỏi dùng để khám phá dữ liệu

1. Đọc dữ liệu và tính số dòng, số cột
2. Mỗi dòng có ý nghĩa gì? Có vấn đề các dòng có ý nghĩa khác nhau không?
3. Dữ liệu có dòng nào bị lặp không?
4. Tỷ lệ giá trị thiếu của mỗi cột
5. Kiểu dữ liệu của mỗi cột
6. Mỗi cột có ý nghĩa gì?
7. Với mỗi cột có dữ liệu số, các giá trị phân bố như thế nào?
8. Với mỗi cột có kiểu dữ liệu không phải dạng số, các giá trị phân bố như thế nào?

# B. Khám phá dữ liệu

## II. Trả lời các câu hỏi khám phá dữ liệu

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

- Import các thư viện cần thiết: pandas, numpy, matplotlib, ...
- Đọc file csv đã được tải về bằng `pandas.read_csv()` và lưu vào biến `df`
- Tính số dòng và số cột thông qua `df.shape`

Unnamed: 0	index	Job Title	Salary Estimate	Job Description	Rating	Company Name	Location	Headquarters	Size	Founded	Type of ownership	Industry	Sector	Revenue	Competitors	Easy Apply	
0	0	0	Senior Data Scientist	\$111K-\$181K (Glassdoor est.)	ABOUT HOPPER\n\nAt Hopper, we're on a mission ...	3.5	Hopper\n3.5	New York, NY	Montreal, Canada	501 to 1000 employees	2007	Company - Private	Travel Agencies	Travel & Tourism	Unknown / Non-Applicable	-1	-1
1	1	1	Data Scientist, Product Analytics	\$111K-\$181K (Glassdoor est.)	At Noom, we use scientifically proven methods ...	4.5	Noom US\n4.5	New York, NY	New York, NY	1001 to 5000 employees	2008	Company - Private	Health, Beauty, & Fitness	Consumer Services	Unknown / Non-Applicable	-1	-1
2	2	2	Data Science Manager	\$111K-\$181K (Glassdoor est.)	Decode_M\n\nhttps://www.decode-m.com/\n\nData ...	-1.0	Decode_M	New York, NY	New York, NY	1 to 50 employees	-1	Unknown	-1	-1	Unknown / Non-Applicable	-1	True
3	3	3	Data Analyst	\$111K-\$181K (Glassdoor est.)	Sapphire Digital seeks a dynamic and driven mi...	3.4	Sapphire Digital\n3.4	Lyndhurst, NJ	Lyndhurst, NJ	201 to 500 employees	2019	Company - Private	Internet	Information Technology	Unknown / Non-Applicable	Zocdoc, Healthgrades	-1
4	4	4	Director, Data Science	\$111K-\$181K (Glassdoor est.)	Director, Data Science - (200537)\nDescription...	3.4	United Entertainment Group\n3.4	New York, NY	New York, NY	51 to 200 employees	2007	Company - Private	Advertising & Marketing	Business Services	Unknown / Non-Applicable	BBDO, Grey Group, Droga5	-1

# B. Khám phá dữ liệu

## II. Trả lời các câu hỏi khám phá dữ liệu

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

- Mỗi dòng trong tập dữ liệu là thông tin về vị trí tuyển dụng liên quan đến ngành Khoa học dữ liệu. Có vẻ không có vấn đề các dòng có ý nghĩa khác nhau, tức là không có dòng nào bị 'lạc loài'

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

- Ta kiểm tra xem có dòng nào bị lặp không bằng cách sử dụng `uplicated()` và `any()` trên dataframe `df` và lưu kết quả vào biến `have_duplicated_rows`. Biến này sẽ có giá trị `True` nếu dữ liệu có các dòng bị lặp và có giá trị `False` nếu không có dòng bị lặp.
- Kết quả sau khi kiểm tra trên `df` → `False` (không có dòng nào bị lặp)

# B. Khám phá dữ liệu

## II. Trả lời các câu hỏi khám phá dữ liệu

### 4. Tỷ lệ giá trị thiếu của mỗi cột?

- Thay thế các giá trị không hợp lệ trong cột `Easy Apply` thành giá trị boolean `False` bằng cách sử dụng `fillna()` và `astype()`
- Thay các giá trị không hợp lệ (bằng -1) khác trong `df` thành `NaN` bằng `replace()`
- Ta cần xem xét đến các giá trị thiếu trong dữ liệu.
  - Tính tỷ lệ giá trị thiếu của từng cột bằng cách sử dụng phương thức `isnull()` trên dataframe `df` và tính tổng số giá trị thiếu của từng cột bằng phương thức `sum()`.
  - Cuối cùng ta chia số dòng và lưu kết quả vào `missing_ratio`.
- Ta cũng cần tính các giá trị thống kê mô tả của các cột numeric bằng phương thức `describe()` trên dataframe `df`.



```

Unnamed: 0      0.000000
index           0.000000
Job Title       0.000000
Salary Estimate 0.000000
Job Description 0.000000
Rating          0.104630
Company Name    0.000000
Location        0.000000
Headquarters    0.061397
Size            0.058583
Founded         0.249936
Type of ownership 0.058583
Industry        0.139678
Sector          0.139678
Revenue         0.058583
Competitors     0.706063
Easy Apply      0.000000
dtype: float64

```

missing\_ratio

	Unnamed: 0	index	Rating	Founded
count	3909.000000	3909.000000	3500.000000	2932.000000
mean	1954.000000	2167.446662	3.784143	1972.371419
std	1128.575429	1247.657849	0.614619	52.719618
min	0.000000	0.000000	1.000000	1625.000000
25%	977.000000	1121.000000	3.400000	1961.000000
50%	1954.000000	2161.000000	3.800000	1995.000000
75%	2931.000000	3249.000000	4.100000	2006.000000
max	3908.000000	4379.000000	5.000000	2020.000000

describe()

# B. Khám phá dữ liệu

## II. Trả lời các câu hỏi khám phá dữ liệu

5. Kiểu dữ liệu của mỗi cột? Có cột nào có kiểu dữ liệu chưa phù hợp để xử lý tiếp không?
- Sử dụng `df.dtypes` để kiểm tra dữ liệu mỗi cột

```
Unnamed: 0      int64
index           int64
Job Title       object
Salary Estimate object
Job Description object
Rating         float64
Company Name    object
Location        object
Headquarters    object
Size            object
Founded         float64
Type of ownership object
Industry        object
Sector          object
Revenue         object
Competitors     object
Easy Apply      bool
dtype: object
```

# B. Khám phá dữ liệu



## II. Trả lời các câu hỏi khám phá dữ liệu

### 6. Mỗi cột có ý nghĩa gì?

- Ý nghĩa của các cột đã được đề cập ở phần Thu thập dữ liệu
- Cột `Unnamed: 0` có giá trị trùng với index → không cần thiết → dùng `drop()` để bỏ

# B. Khám phá dữ liệu

## II. Trả lời các câu hỏi khám phá dữ liệu

### 7. Với mỗi cột có kiểu dữ liệu số, các giá trị phân bố như thế nào?

- Với các cột có kiểu dữ liệu số, ta sẽ tính:
  - Tỷ lệ % (từ 0 đến 100) các giá trị thiếu
  - Giá trị min
  - Giá trị lower quartile (phần vị 25)
  - Giá trị median (phần vị 50)
  - Giá trị upper quartile (phần vị 75)
  - Giá trị max
- Lưu kết quả vào DataFrame `num_col_info_df`, trong đó:
  - Tên của các cột là tên của các cột số trong `df`
  - Tên của các dòng là: `missing_ratio`, `min`, `lower_quartile`, `median`, `upper_quartile`, `max`



	index	Rating	Founded
row_name			
missing_ratio	0.0	10.5	25.0
min	0.0	1.0	1625.0
lower_quartile	1121.0	3.4	1961.0
median	2161.0	3.8	1995.0
upper_quartile	3249.0	4.1	2006.0
max	4379.0	5.0	2020.0

# B. Khám phá dữ liệu

## II. Trả lời các câu hỏi khám phá dữ liệu

8. Với mỗi cột có kiểu dữ liệu không phải dạng số, các giá trị được phân bố như thế nào?

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

- Tỷ 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`).
- Tỷ lệ % (từ 0 đến 100) của mỗi giá trị được sort theo tỷ lệ % giảm dần (không xét giá trị thiếu, tỷ lệ là tỷ lệ so với số lượng các giá trị không thiếu): dùng 1 dictionary để lưu, key là giá trị, value là tỷ lệ % (`value_ratios`).

	Job Title	Job Description	Company Name	Type of ownership	Industry	Sector	Revenue	Competitors	Location City	Location State	Headquarters City	Headquarters State
row_name												
missing_ratio	0.0	0.0	0.0	5.858276	13.967767	13.967767	5.858276	70.606293	0.0	0.0	6.139678	6.139678
num_values	2079	3685	2069	14	95	25	13	421	185	11	523	80
value_ratios	{'Data Scientist': 7.0, 'Data Engineer': 6.7, ...	{'The U.S. Department of the Treasury has a di...	{'Apple ': 1.5, 'IBM ': 1.5, 'Amazon ': 1.2, ' ...	{'Company - Private': 52.5, 'Company - Public'...	{'IT Services': 14.0, 'Staffing & Outsourcing'...	{'Information Technology': 33.4, 'Business Ser...	{'Unknown / Non-Applicable': 31.6, '\$10+ billi...	{'Amazon, Accenture, Microsoft': 5.3, 'Google,...	{'Austin': 8.8, 'Chicago': 8.4, 'San Diego': 7...	{' TX': 32.1, ' CA': 27.3, ' IL': 9.3, ' PA': ...	{'New York': 8.2, 'San Diego': 4.9, 'Chicago':...	{' CA': 24.9, ' TX': 12.5, ' NY': 10.6, ' IL':...

Unnamed: 0	index	Job Title	Salary Estimate	Job Description	Rating	Company Name	Location	Headquarters	Size	Founded	Type of ownership	Industry	Sector	Revenue	Competitors	Easy Apply	
0	0	0	Senior Data Scientist	\$111K-\$181K (Glassdoor est.)	ABOUT HOPPER\n\nAt Hopper, we're on a mission ...	3.5	Hopper\n3.5	New York, NY	Montreal, Canada	501 to 1000 employees	2007	Company - Private	Travel Agencies	Travel & Tourism	Unknown / Non-Applicable	-1	-1
1	1	1	Data Scientist, Product Analytics	\$111K-\$181K (Glassdoor est.)	At Noom, we use scientifically proven methods ...	4.5	Noom US\n4.5	New York, NY	New York, NY	1001 to 5000 employees	2008	Company - Private	Health, Beauty, & Fitness	Consumer Services	Unknown / Non-Applicable	-1	-1
2	2	2	Data Science Manager	\$111K-\$181K (Glassdoor est.)	Decode_M\n\nhttps://www.decode-m.com/\n\nData ...	-1.0	Decode_M	New York, NY	New York, NY	1 to 50 employees	-1	Unknown	-1	-1	Unknown / Non-Applicable	-1	True
3	3	3	Data Analyst	\$111K-\$181K (Glassdoor est.)	Sapphire Digital seeks a dynamic and driven mi...	3.4	Sapphire Digital\n3.4	Lyndhurst, NJ	Lyndhurst, NJ	201 to 500 employees	2019	Company - Private	Internet	Information Technology	Unknown / Non-Applicable	Zocdoc, Healthgrades	-1
4	4	4	Director, Data Science	\$111K-\$181K (Glassdoor est.)	Director, Data Science - (200537)\n\nDescription...	3.4	United Entertainment Group\n3.4	New York, NY	New York, NY	51 to 200 employees	2007	Company - Private	Advertising & Marketing	Business Services	Unknown / Non-Applicable	BBDO, Grey Group, Droga5	-1

Dữ liệu trước khi khám phá và tiền xử lý

index		Job Title	Salary Estimate	Job Description	Rating	Company Name	Location	Headquarters	Size	Founded	Type of ownership	Industry	Sector	Revenue	Competitors	Easy Apply
0	0	Senior Data Scientist	\$111K-\$181K (Glassdoor est.)	ABOUT HOPPER\n\nAt Hopper, we're on a mission ...	3.5	Hopper\n3.5	New York, NY	Montreal, Canada	501 to 1000 employees	2007.0	Company - Private	Travel Agencies	Travel & Tourism	Unknown / Non-Applicable	NaN	True
1	1	Data Scientist, Product Analytics	\$111K-\$181K (Glassdoor est.)	At Noom, we use scientifically proven methods ...	4.5	Noom US\n4.5	New York, NY	New York, NY	1001 to 5000 employees	2008.0	Company - Private	Health, Beauty, & Fitness	Consumer Services	Unknown / Non-Applicable	NaN	True
2	2	Data Science Manager	\$111K-\$181K (Glassdoor est.)	Decode_M\n\nhttps://www.decode-m.com/\n\nData ...	NaN	Decode_M	New York, NY	New York, NY	1 to 50 employees	NaN	Unknown	NaN	NaN	Unknown / Non-Applicable	NaN	True
3	3	Data Analyst	\$111K-\$181K (Glassdoor est.)	Sapphire Digital seeks a dynamic and driven mi...	3.4	Sapphire Digital\n3.4	Lyndhurst, NJ	Lyndhurst, NJ	201 to 500 employees	2019.0	Company - Private	Internet	Information Technology	Unknown / Non-Applicable	Zocdoc, Healthgrades	True
4	4	Director, Data Science	\$111K-\$181K (Glassdoor est.)	Director, Data Science - (200537)\n\nDescription...	3.4	United Entertainment Group\n3.4	New York, NY	New York, NY	51 to 200 employees	2007.0	Company - Private	Advertising & Marketing	Business Services	Unknown / Non-Applicable	BBDO, Grey Group, Droga5	True

Dữ liệu sau khi khám phá và tiền xử lý



# C&D. Đặt và phân tích trả lời câu hỏi

## Câu hỏi 1:

### 1. Nội dung câu hỏi

Với mỗi công việc được tuyển thì nên chọn công ty nào là nơi làm việc tốt nhất?

### 2. Lợi ích khi trả lời câu hỏi

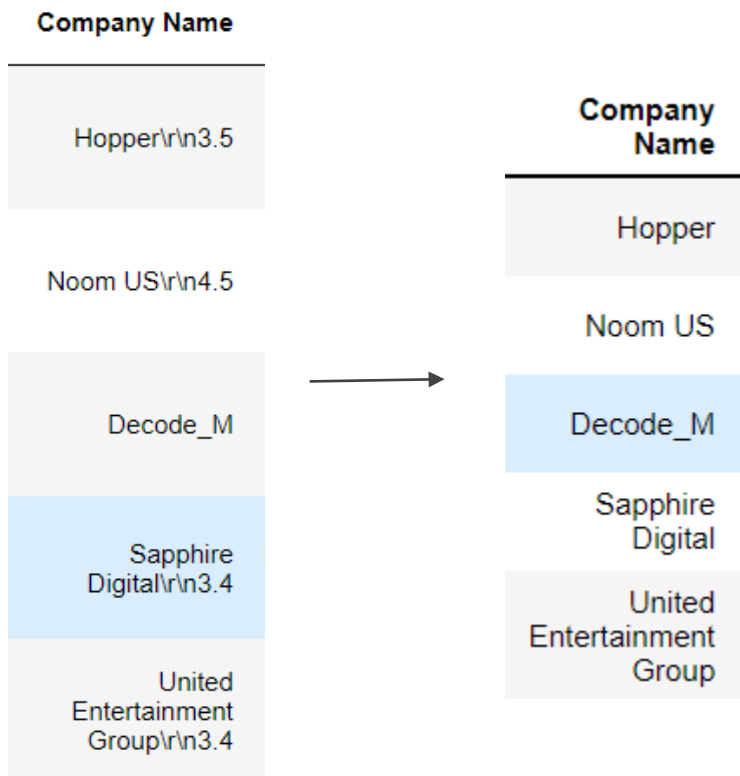
Mỗi vị trí công việc có nhiều công ty để lựa chọn. Việc lọc như thế này có thể đưa ra sự lựa chọn công ty tốt nhất cho mỗi vị trí công việc.

### 3. Tiềm xử lý, phân tích để trả lời câu hỏi

- Xử lý dữ liệu cột **Company Name**: xóa bớt các kí tự thừa ('/r/n')
- Xử lý dữ liệu cột **Salary Estimate**: tách ra hai cột mới là **Minimum Salary** và **Maximum Salary**; tính trung bình và tạo ra cột **Avg Salary**
- Gom nhóm dữ liệu trong cột **Job Title** để chọn **Company Name** tốt nhất theo tiêu chí **Rating** và **Avg Salary** cao nhất
- Đếm số lượng **Company Name** có trong kết quả trên và cho ra một vài nhận xét.

### 3. Tiền xử lý, phân tích để trả lời câu hỏi

- Xử lý dữ liệu cột **Company Name**: xóa bớt các kí tự thừa (/r/n)  
Dùng hàm **replace()** để xóa phần dữ liệu thừa này



The diagram illustrates the transformation of a 'Company Name' column. On the left, the original data contains carriage returns and newlines. An arrow points to the right, showing the result after applying the `replace()` function to remove these characters.

Company Name		Company Name
Hopper\r\n3.5		Hopper
Noom US\r\n4.5		Noom US
Decode_M		Decode_M
Sapphire Digital\r\n3.4		Sapphire Digital
United Entertainment Group\r\n3.4		United Entertainment Group

### 3. Tiền xử lý, phân tích để trả lời câu hỏi

- Xử lý dữ liệu cột **Salary Estimate**: tách ra hai cột mới là **Minimum Salary** và **Maximum Salary**; tính trung bình và tạo ra cột **Avg Salary**  
Dùng hàm **replace()** để xóa bớt kí tự thừa và hàm **split()** để lấy ra 2 giá trị tiền lương đó, rồi dùng hàm **astype()** để chuyển từ kiểu chuỗi sang dạng numeric.

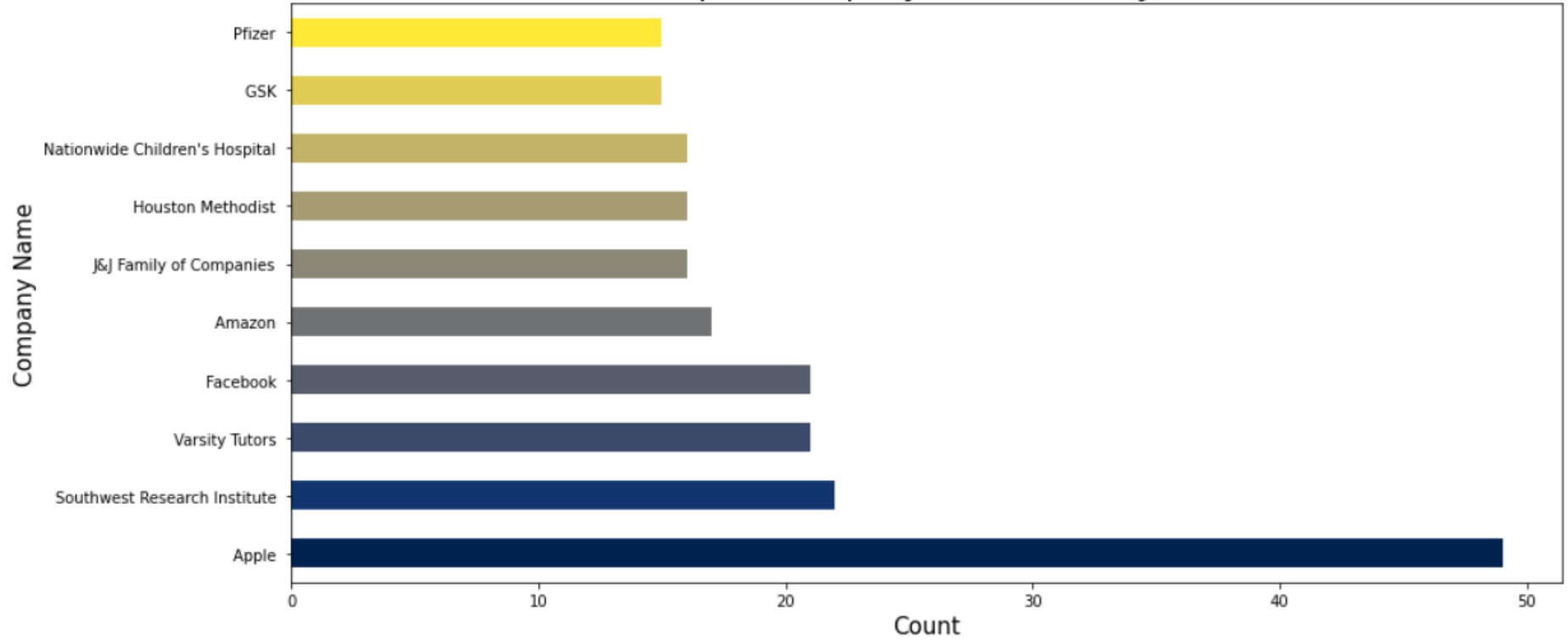
Salary Estimate	Minimum Salary	Maximum Salary	Avg Salary
111K–181K (Glassdoor est.)	111	181	146.0
111K–181K (Glassdoor est.)	111	181	146.0
111K–181K (Glassdoor est.)	111	181	146.0



### 3. Tiền xử lí, phân tích để trả lời câu hỏi

- Gom nhóm dữ liệu trong cột Job Title để chọn Company Name tốt nhất theo tiêu chí Rating và Avg Salary cao nhất.
  - Ta sử dụng hàm `groupby()` để gom nhóm dữ liệu trong cột Job Title
  - Kết hợp với hàm `max()` để lọc ra Company Name có Rating và Avg Salary cao nhất.
- Đếm số lượng Company Name trong kết quả trên, trực quan hóa và cho ra một vài nhận xét
  - Ta dùng hàm `value_counts()` để đếm số giá trị Company Name trong dataframe kết quả trên.
  - Sử dụng biểu đồ `plot.barh()` để biểu diễn top các công ty được đánh giá tốt cho các công việc và cho ra nhận xét.

Top 10 Company best for Data Job



# C&D. Đặt và phân tích trả lời câu hỏi

## Câu hỏi 2:

### 1. Nội dung câu hỏi:

Lĩnh vực có doanh thu cao nhất và số lượng nhân viên trong từng lĩnh vực có liên quan gì đến nhau?

### 2. Lợi ích khi trả lời câu hỏi:

- Có thể trực quan hóa tổng doanh thu của từng lĩnh vực, biết được xếp hạng doanh thu của từng lĩnh vực.
- Kết hợp quy mô lực lượng và doanh thu thì ta có thể biết được sự tương quan giữa chúng.

### 3. Tiền xử lý, phân tích để trả lời câu hỏi:

- Xử lý dữ liệu trong cột doanh thu (**Revenue**): xử lý đơn vị của doanh thu (million, billion,...), xử lý dữ liệu số trong cột ('to', 'less than',...)
- Liệt kê danh sách các lĩnh vực.
- Tính số lượng nhân viên trung bình (**Size**), doanh thu (**Revenue**) trung bình của từng lĩnh vực.
- Vẽ biểu đồ trực quan hóa dữ liệu và từ đó đưa ra nhận xét, câu trả lời.

### 3. Tiền xử lý, phân tích để trả lời câu hỏi:

- Xử lý dữ liệu trong cột doanh thu (Revenue):
  - + Đầu tiên chúng ta sẽ xóa đi các dữ liệu trống và xóa các ký tự như '\$','USD',...có trong dữ liệu bằng cách sử dụng:  
`df[column_name].replace('$','')`
  - + Sau bước trên, dữ liệu trong cột Revenue sẽ còn lại 2 loại:

1. 'AtoB': sử dụng hàm `split('to')` để chia dữ liệu ra thành 2 cột **Minimum** và **Maximum**

2. 'lessthanN' / 'N+': trước tiên ta sẽ thay thế 'lessthan' và '+' bằng cách sử dụng `replace()` để chuyển nó về định dạng 'AtoB' như loại 1 và xử lý nó.

3. Tính doanh thu trung bình của 2 cột **Min** và **Max** sau đó lưu vào cột **AVG Revenue**

- Đối với xử lý cột **Size** thì chúng ta cũng xử lý tương tự như cột **Revenue**

Revenue (	Minimum Revenue	Maximum Revenue	Size	Minimum Size	Maximum Size
10to25 million (USD)	10.0	25.0	201 to 500 employees	201.0	500.0
50to100 million (USD)	50.0	100.0	51 to 200 employees	51.0	200.0
100to 500 million (USD)	100.0	500.0	5001 to 10000 employees	5001.0	10000.0
Less than \$1 million (USD)	0.1	1.0	1 to 50 employees	1.0	50.0
2to5 billion (USD)	2000.0	5000.0	5001 to 10000 employees	5001.0	10000.0

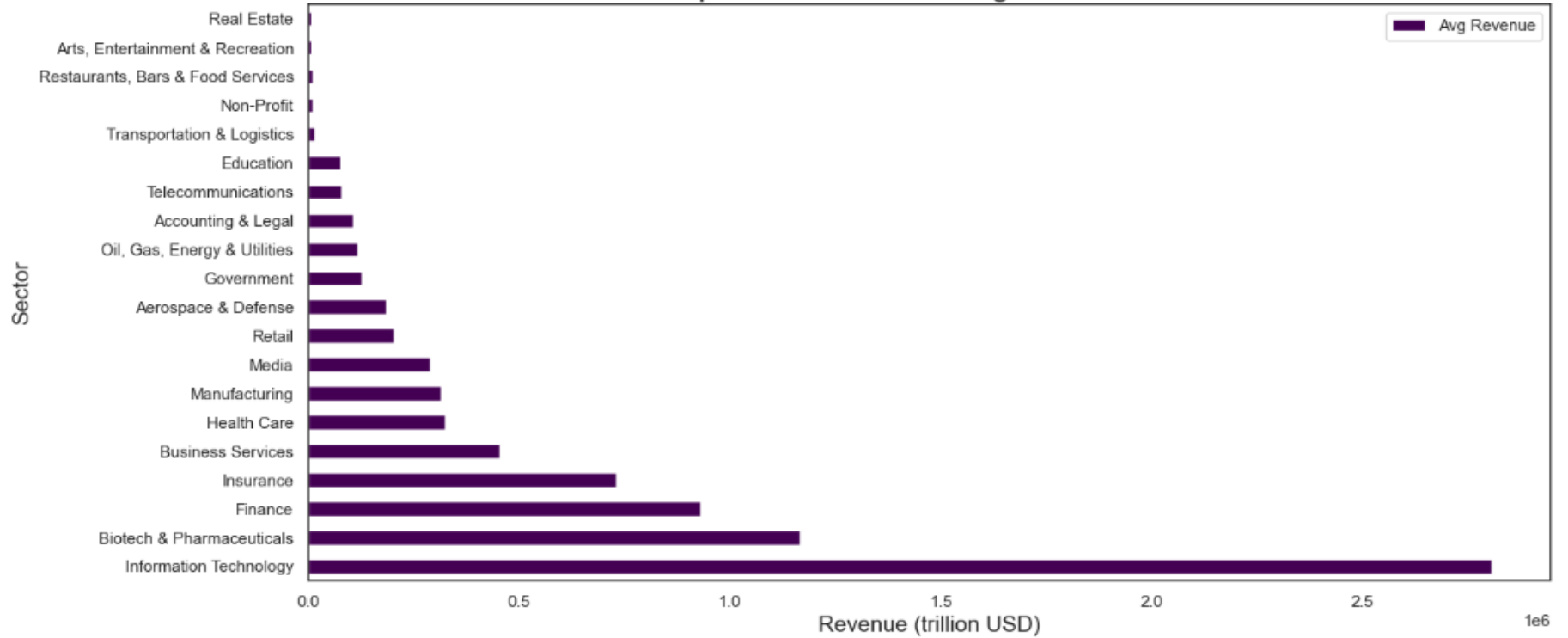




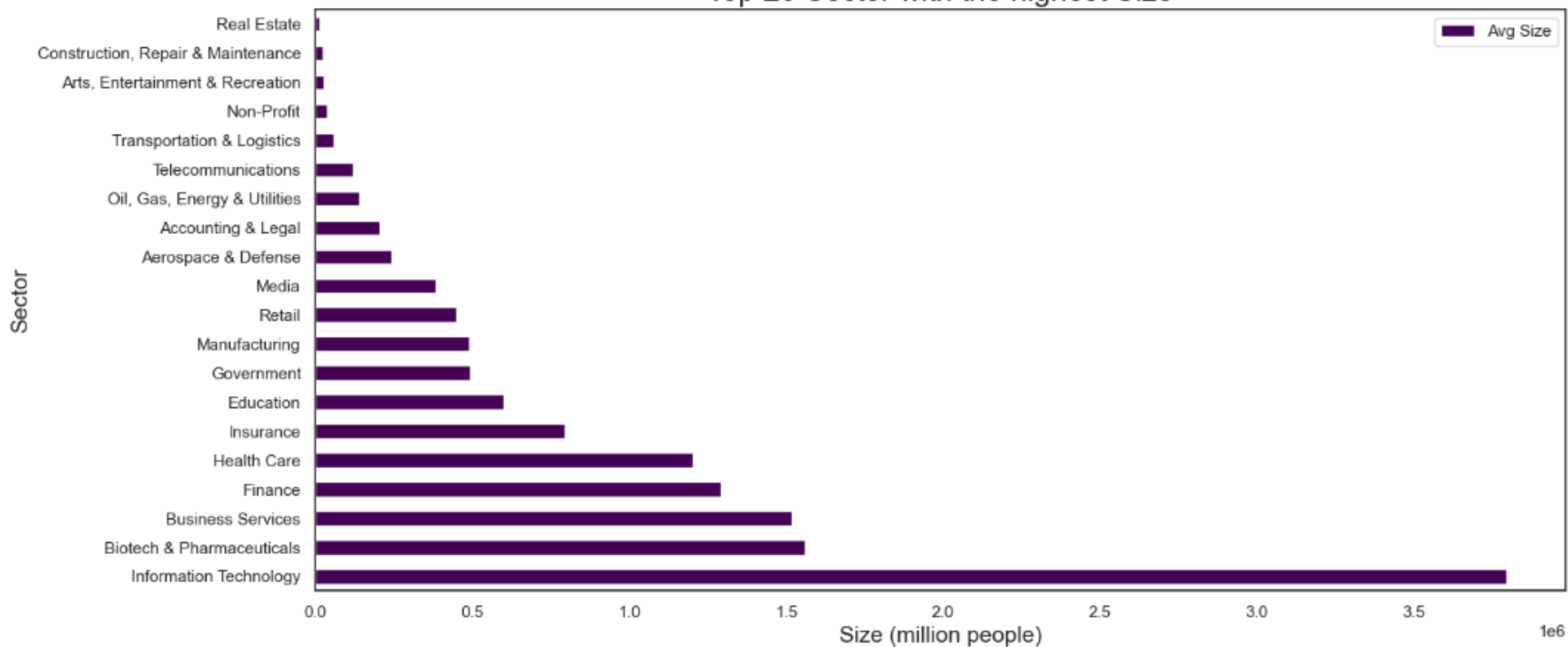
### 3. Tiền xử lý, phân tích để trả lời câu hỏi:

- Vẽ biểu đồ trực quan hóa dữ liệu và từ đó đưa ra nhận xét, câu trả lời:
  1. Đối với câu hỏi so sánh dữ liệu thì sử dụng `plot.barh()` để tiện cho việc quan sát và đánh giá dữ liệu giữa từng cột.
  2. Lấy top 20 lĩnh vực dẫn đầu trong mỗi mục đích so sánh để vẽ biểu đồ `.head(20)`
  3. Quan sát biểu đồ và tự đưa ra nhận xét của bản thân.

Top 20 Sector with the highest Revenue



Top 20 Sector with the highest Size



# C&D. Đặt và phân tích trả lời câu hỏi

## Câu hỏi 3:

### 1. Nội dung câu hỏi

Xu hướng tuyển dụng của các công ty lớn hiện nay như thế nào?

Kèm theo các câu hỏi nhỏ bao gồm:

Các công ty lớn hiện nay là những công ty nào?

Các công việc được các công ty tuyển dụng nhiều nhất?

### 2. Lợi ích khi trả lời câu hỏi

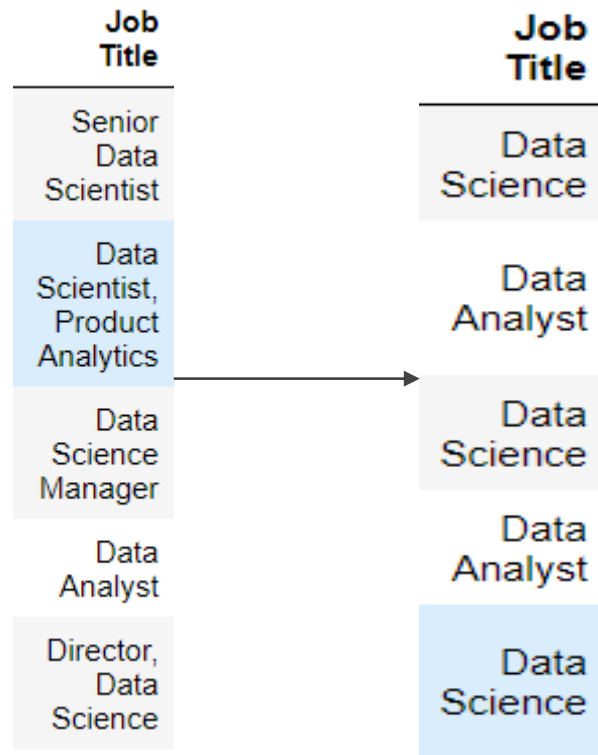
Ta có thể biết được các vị trí tuyển dụng đang bị thiếu hụt nhiều nhất trên thị trường nhân sự

- Đối với doanh nghiệp -> nhanh chóng nắm bắt được xu hướng kinh doanh và phát triển của công ty khác
- Đối với các nhà môi giới tuyển dụng -> dễ dàng nắm bắt và nhanh chóng tìm kiếm nguồn nhân sự cho các công ty
- Đối với người tìm việc -> kịp thời nắm bắt xu hướng việc làm để phát triển bản thân và tìm kiếm công việc phù hợp

# C&D. Đặt và phân tích trả lời câu hỏi

## Câu hỏi 3:

3. Tiền xử lý, phân tích để trả lời câu hỏi:
- Lọc lại **Job Title** bằng từ khóa (do **Job Title** không theo một form nhất định) bằng cách thay các toàn bộ ô job title bằng các nghề chung theo các từ khóa đó: Ta sử dụng hàm **replace()**



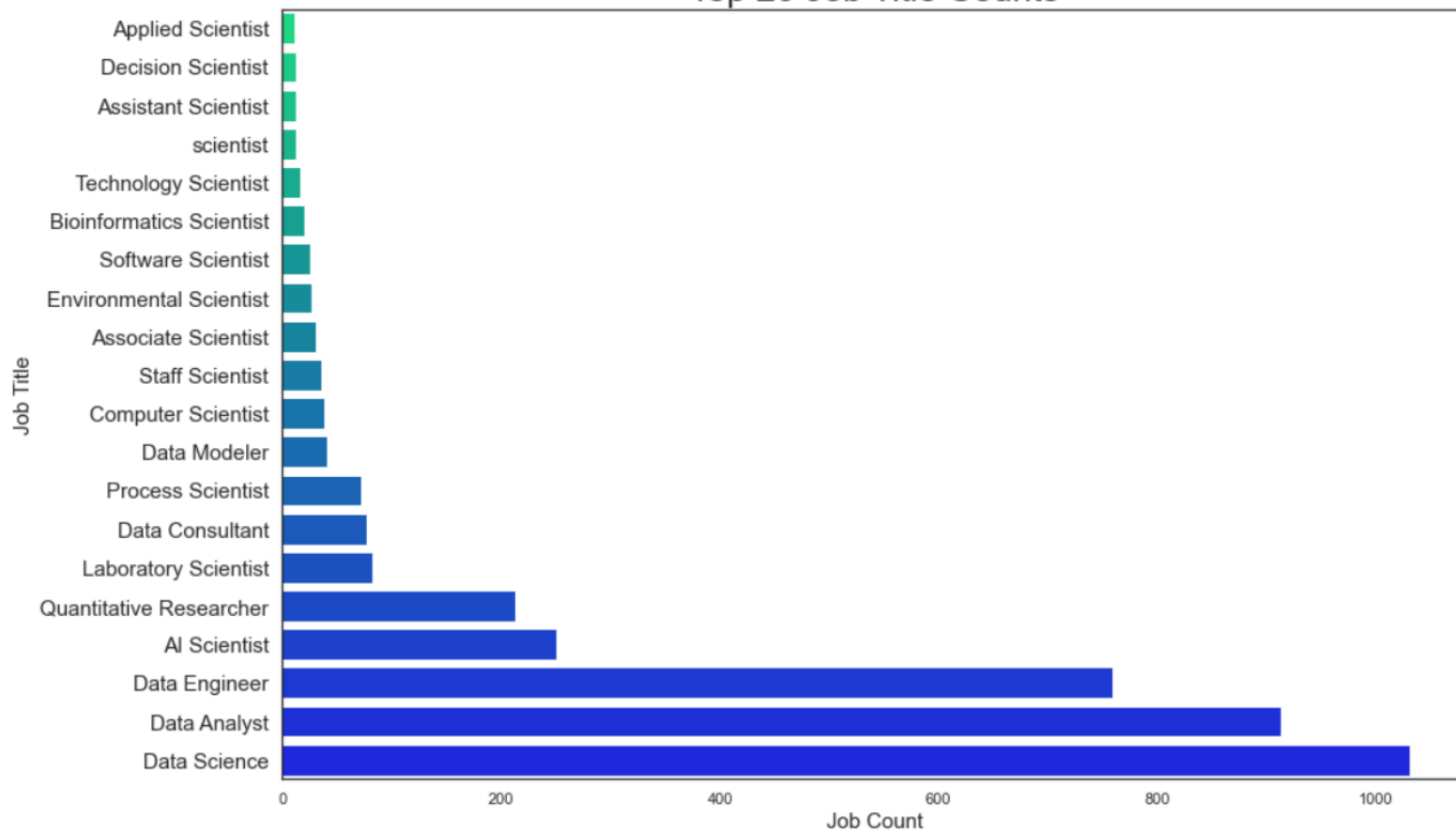
# C&D. Đặt và phân tích trả lời câu hỏi

## Câu hỏi 3:

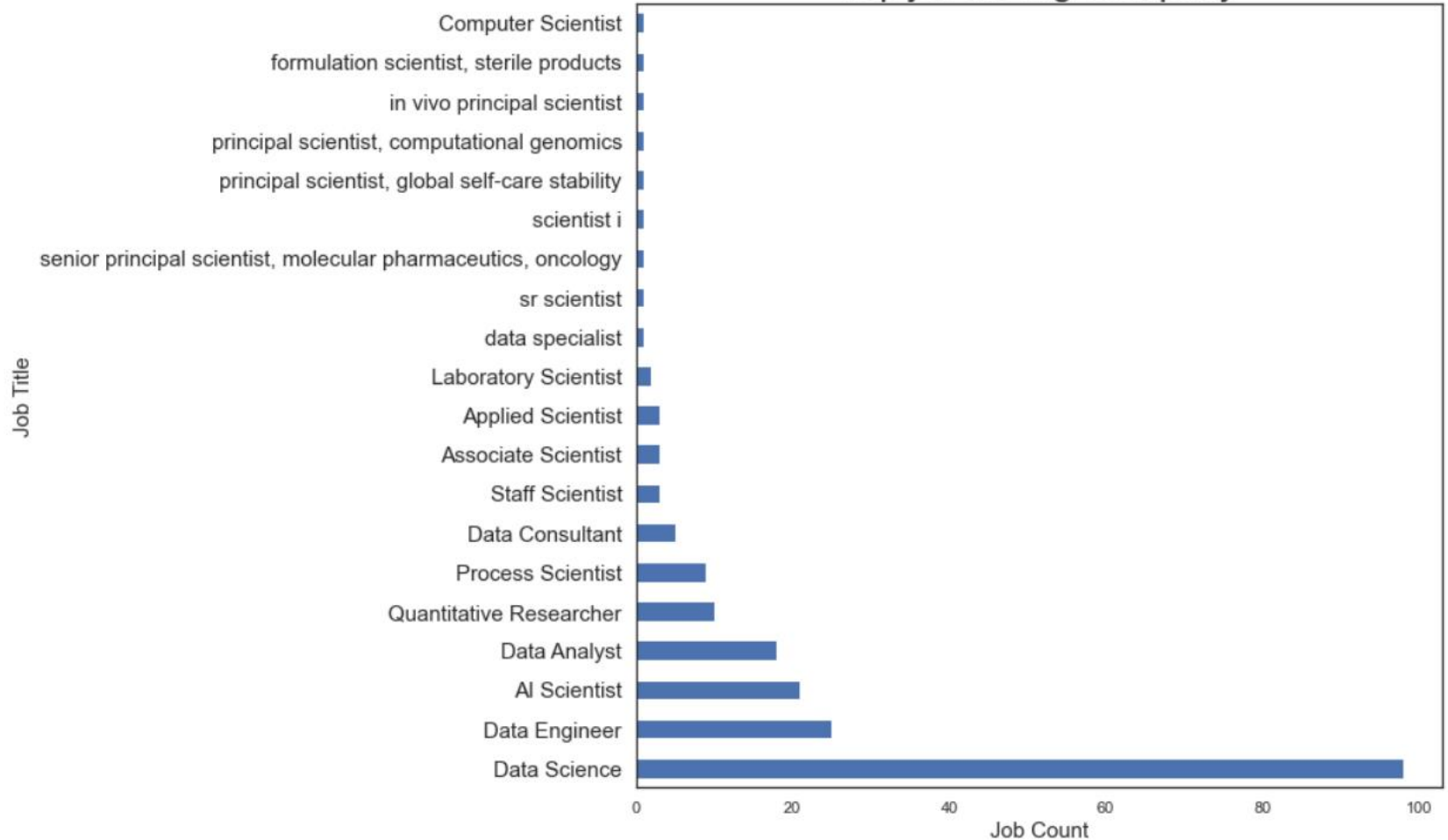
### 3. Tiền xử lý, phân tích để trả lời câu hỏi

- Tiêu chí đánh giá các công ty lớn: **Rating**  $\geq 3$ , **Revenue** min  $> 10000$ , **Size** min  $> 10000$ , **Salary** min  $> 10000$ .
- Thực hiện vẽ biểu đồ cột: các Job được tuyển dụng nhiều nhất và các Job được các công ty lớn tuyển dụng nhiều nhất

Top 20 Job Title Counts



## Top job for Big Company





# C&D. Đặt và phân tích trả lời câu hỏi



## Câu hỏi 4:

### 1. Nội dung câu hỏi

Đâu là nơi tuyển dụng nhân lực Khoa học dữ liệu nhiều nhất?

### 2. Lợi ích khi trả lời câu hỏi

Biết được đâu là nơi tuyển dụng về lĩnh vực Khoa học dữ liệu nhiều nhất, suy ra được đâu là nơi mà Khoa học dữ liệu đang phát triển mạnh nhất

# C&D. Đặt và phân tích trả lời câu hỏi

## Câu hỏi 4:

1. Nội dung câu hỏi
2. Lợi ích khi trả lời câu hỏi
3. Tiền xử lý, phân tích để trả lời câu hỏi
  - Tách cột `Location` và `Headquarters` thành `Location City`, `Location State`, `Headquarters City` và `Headquarters State`
  - Đếm số lượng thành phố đang cần tuyển nhân viên bằng `values_counts()`, sau đó lấy top 20. Ý nghĩa của các cột đã được đề cập ở phần Thu thập dữ liệu thành phố đầu tiên và lưu vào `df_by_city`
  - Gộp 2 dataframe `df_by_city` và `df` bằng `merge()` với từ khóa `Location City`. Sau đó lưu vào `Sal_by_city`
  - Dùng `plt.subplots()` để vẽ biểu đồ cột top 20 thành phố tuyển dụng nhiều nhất và biểu đồ pointplot thể hiện lương trung bình tương ứng với mỗi thành phố

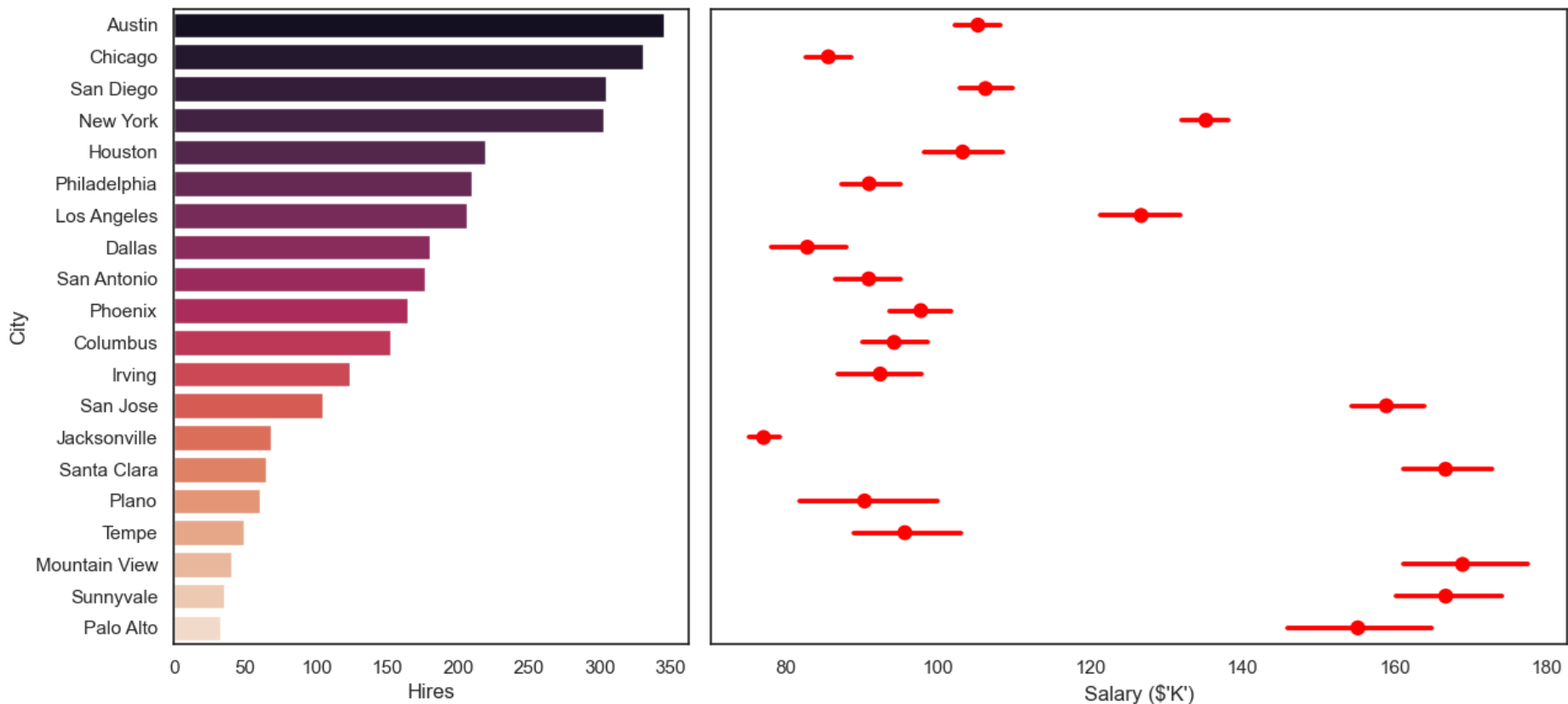
Cột **Location** trước và sau khi xử lý

<b>Location</b>	<b>Location City</b>	<b>Location State</b>
New York, NY	New York	NY
New York, NY	New York	NY
New York, NY	New York	NY
Lyndhurst, NJ	Lyndhurst	NJ
New York, NY	New York	NY

Cột **Headquarters** trước và sau khi xử lý

<b>Headquarters</b>	<b>Headquarters City</b>	<b>Headquarters State</b>
Montreal, Canada	Montreal	Canada
New York, NY	New York	NY
New York, NY	New York	NY
Lyndhurst, NJ	Lyndhurst	NJ
New York, NY	New York	NY

## Top 20 Cities Hiring Data Science Jobs



# E. Tổng hợp lại quá trình làm đồ án



## Khó khăn:

- Khai thác dữ liệu để đặt câu hỏi
- Xử lý dữ liệu gặp nhiều khó khăn, có những giá trị xử lý chưa thực sự hiệu quả trong quá trình làm
- Lần đầu sử dụng github
- Sử dụng biểu đồ

## Học được:

- Làm việc nhóm
- Sử dụng Trello để giao và quản lý công việc
- Sử dụng git và github
- Xử lý dữ liệu
- Các hàm, thư viện mới, dạng biểu đồ mới, cách vẽ mới

# E. Tổng hợp lại quá trình làm đồ án



## Nếu có thêm nhiều thời gian hơn:

- Tìm cách xử lý phù hợp hơn đối với một số kiểu dữ liệu của cột
- Cố gắng tìm những nguồn dữ liệu hay hơn và có thể sẽ tìm cách tự crawl dữ liệu về
- Tìm hiểu thêm để có thể đặt được những câu hỏi hay và ý nghĩa hơn
- Tìm hiểu thêm về các loại biểu đồ và ứng dụng của chúng, cũng như tìm hiểu thêm cách để biểu diễn biểu đồ đẹp mắt và truyền đạt nhiều thông tin hơn
- Tìm hiểu thêm về cơ chế pull request của github



# TÀI LIỆU THAM KHẢO

- <https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.replace.html>
- <http://thongthai.work/cach-thuc-hien-phan-tich-du-lieu-kham-pha/>
- <https://matplotlib.org/stable/tutorials/colors/colormaps.html>