ĐẠI HỌC HUẾ
KHOA KỸ THUẬT VÀ CÔNG NGHỆ

BÁO CÁO ĐỒ ÁN
HỌC KÌ II, năm học 2022-2023

Học phần:
PHÂN TÍCH DỮ LIỆU VỚI PYTHON

Số phách **(**Do hội đồng chấm ghi thi)

Thừa Thiên Huế, tháng 04 năm 2023.

ĐẠI HỌC HUẾ
KHOA KỸ THUẬT VÀ CÔNG NGHỆ

BÁO CÁO ĐỒ ÁN
HỌC KÌ II, năm học 2022-2023

Học phần:
PHÂN TÍCH DỮ LIỆU VỚI PYTHON

Giảng viên hướng dẫn: Nguyễn Thế Dũng
Lớp: Khoa học dữ liệu và Trí tuệ nhân tạo khóa 3
Sinh viên thực hiện: Nhóm 5_Phạm Phước Bảo Tín_22E1020021
(ký và ghi rõ họ tên)

Số phách **(**Do hội đồng chấm ghi thi)

Thừa Thiên Huế, tháng 04 năm 2023

ĐẠI HỌC HUẾ
KHOA KỸ THUẬT VÀ CÔNG NGHỆ

PHIẾU ĐÁNH GIÁ ĐỒ ÁN/TIỂU LUẬN/BÀI TẬP LỚN
Học kỳ II, năm học 2022 - 2023

Cán bộ chấm thi 1	**Cán bộ chấm thi 2**
Nhận xét: .... **Điểm đánh giá của CBCT1:** Bằng số: ..... Bằng chữ:	**Nhận xét:** .... **Điểm đánh giá của CBCT2:** Bằng số: ..... Bằng chữ: .....

Điểm kết luận: .....

Bằng số:.....

Bằng chữ: .....

Thừa Thiên Huế, ngày 07 tháng 06 năm 2023

Cán bộ chấm thi 1	**Cán bộ chấm thi 2**
(Ký và ghi rõ họ và tên)	*(Ký và ghi rõ họ và tên)*

LỜI CẢM ƠN

Được trở thành sinh viên Khoa Kỹ Thuật và Công Nghệ - Đại học Huế em rất hạnh phúc và biết ơn. Hạnh phúc vì mình đã đạt được mục tiêu mong muốn và biết ơn sự cống hiến, chỉ bảo tận tình sâu sắc của quý thầy cô trong khoa đồng thời đã tạo điều kiện học tập lí tưởng cho chúng em. Để hoàn thành đồ án một cách chỉnh chu nhất có thể em xin gửi lời cảm ơn đến thầy giáo bộ môn – Thầy Nguyễn Thế Dũng đã hướng dẫn tận tình, chi tiết cho chúng em trong quá trình hoàn thành đồ án lẫn quá trình học tập học của sinh viên chúng em. Hi vọng rằng thời gian sắp tới em sẽ luôn cố gắng, nổ lực hơn nữa trong học tập chuyên nghành của mình.

Trong quá trình hoàn thành đồ án mặc dù em đã chuẩn bị kĩ nhưng không thể tránh khỏi những sai sót, em mong nhận được sự góp ý từ quý thầy, cô. Lời cuối cùng em xin kính chúc quý thầy, cô thật nhiều sức khỏe để tiếp tục dẫn dắt chúng em và những thế hệ tiếp theo thành người.

DANH MỤC HÌNH ẢNH

NHIỆM VỤ MỖI THÀNH VIÊN

Sau khi hội nhóm để phân công nhiệm vụ hoàn thành đồ án, Nhóm 5 xin được trình bày về công việc mỗi thành viên như sau:

1. Trần Tùng Dương: Giới thiệu dữ liệu, mô tả dữ liệu, viết báo cáo đồ án;
2. Trịnh Quốc Dân: Đọc dữ liệu và xử lí dữ liệu, vẽ biểu đồ, bài toán ước lượng;
3. Văn Khiêm Chương: Gộp nhóm, bài toán tương quan;
4. Phạm Phước Bảo Tín: Thống kê cơ bản, vẽ biểu đồ, bài toán kiểm định.

MỤC LỤC

PHẦN 1: KHÁI QUÁT DỮ LIỆU CỦA NHÓM

1.1 Giới thiệu dữ liệu

Trong bài báo cáo này nhóm sử dụng dữ liệu từ file "ds_salaries.xlsx" thu thập trên internet là một tập hợp các thông tin về mức lương của nhân viên trong lĩnh vực khoa học dữ liệu và máy học. File này chứa các thuộc tính như năm làm việc, mức độ kinh nghiệm, loại hình làm việc, chức danh công việc, mức lương, đơn vị tiền tệ, mức lương chuyển đổi sang đơn vị USD, quốc gia cư trú của nhân viên, tỷ lệ làm việc từ xa, địa điểm của công ty và quy mô công ty.

Thông qua dữ liệu này, chúng ta có thể khám phá và phân tích các yếu tố quan trọng liên quan đến mức lương trong lĩnh vực này. Chẳng hạn, chúng ta có thể tìm hiểu sự ảnh hưởng của mức độ kinh nghiệm, chức danh công việc và địa điểm làm việc đến mức lương của nhân viên. Ngoài ra, cũng có thể khám phá sự tương quan giữa mức lương và các yếu tố như loại hình làm việc, quốc gia cư trú, tỷ lệ làm việc từ xa và quy mô công ty.

Dữ liệu trong file "ds_salaries.xlsx" cung cấp thông tin đa dạng và chi tiết, cho phép chúng ta thực hiện các phân tích và so sánh để hiểu rõ hơn về thị trường lương trong ngành khoa học dữ liệu và máy học.

Dữ liệu của nhóm sử dụng: ds_salaries - Google Trang tính

1.2 Mô tả dữ liệu

1.2.1 Thực thể

Trong tập dữ liệu của nhóm có hai thực thể chính được đề cập:

1. Nhân viên (Employee): Đại diện cho các cá nhân làm việc trong lĩnh vực khoa học dữ liệu và máy học. Thông tin về nhân viên bao gồm mức độ kinh nghiệm, loại hình làm việc, chức danh công việc, mức lương, đơn vị tiền tệ, mức lương chuyển đổi sang USD, quốc gia cư trú và tỷ lệ làm việc từ xa.
2. Công ty (Company): Đại diện cho các công ty hoạt động trong lĩnh vực khoa học dữ liệu và máy học. Thông tin về công ty bao gồm địa điểm, quy mô và vị trí của công ty.

Các thuộc tính trong file đề cập đến các thông tin của các thực thể này và cho phép phân tích và so sánh các yếu tố liên quan đến mức lương và các yếu tố khác trong lĩnh vực này.

1.2.2 Thuộc tính

Dữ liệu cần phân tích chứa thông tin về mức lương của các nhân viên trong lĩnh vực khoa học dữ liệu và máy học. Các thuộc tính trong dữ liệu được mô tả như sau:

1. work_year: Năm làm việc, đại diện cho năm mà dữ liệu lương được thu thập.
2. experience_level: Mức độ kinh nghiệm của nhân viên, được biểu thị bằng các mã viết tắt, ví dụ SE (Senior Engineer) hoặc MI (Mid-level).
3. employment_type: Loại hình làm việc, có thể là FT (toàn thời gian) hoặc CT (hợp đồng).
4. job_title: Chức danh công việc của nhân viên, ví dụ Principal Data Scientist, ML Engineer, Applied Scientist, Data Scientist,...

5. salary: Mức lương của nhân viên, được ghi bằng đơn vị tiền tệ tương ứng với thuộc tính salary_currency.
6. salary_currency: Đơn vị tiền tệ được sử dụng để biểu thị mức lương, ví dụ EUR (Euro) hoặc USD (US Dollar).
7. salary_in_usd: Mức lương chuyển đổi sang đơn vị USD (US Dollar), cho phép so sánh trực tiếp các mức lương dựa trên cùng một đơn vị.
8. employee_residence: Quốc gia hoặc khu vực cư trú của nhân viên.
9. remote_ratio: Tỷ lệ làm việc từ xa (remote) của nhân viên, được biểu thị dưới dạng phần trăm (%).
10. company_location: Địa điểm của công ty, tương ứng với quốc gia hoặc khu vực.
11. company_size: Quy mô của công ty, có thể là L (lớn) hoặc M (vừa).

Đây là các thuộc tính mô tả thông tin quan trọng về mức lương, chức danh công việc, đơn vị tiền tệ, kinh nghiệm, quốc gia cư trú và các yếu tố liên quan đến công ty và loại hình làm việc.

PHẦN 2: THAO TÁC VỚI DỮ LIỆU BẰNG PYTHON

2.1 Xử lí dữ liệu

2.1.1 Đọc dữ liệu

Muốn đọc dữ liệu ta có thể sử dụng Python kết hợp thư viện Pandas, có thể đọc toàn bộ dữ liệu hoặc đọc tùy chọn lưu vào một dataframe.

Dưới đây là kết quả đọc toàn bộ dữ liệu như: Hình 1.

	work_year	experience_level	employment_type	...	remote_ratio	company_location	company_size
0	2023	SE	FT	...	100	ES	L
1	2023	MI	CT	...	100	US	S
2	2023	MI	CT	...	100	US	S
3	2023	SE	FT	...	100	CA	M
4	2023	SE	FT	...	100	CA	M
...	...	...	...	...	...	...	...
3750	2020	SE	FT	...	100	US	L
3751	2021	MI	FT	...	100	US	L
3752	2020	EN	FT	...	100	US	S
3753	2020	EN	CT	...	100	US	L
3754	2021	SE	FT	...	50	IN	L

[3755 rows x 11 columns]

Hình 1: Kết quả đọc toàn bộ dữ liệu

Dưới đây là kết quả đọc độ lớn tùy chọn của bộ dữ liệu như:

	work_year	experience_level	employment_type	...	remote_ratio	company_location	company_size
0	2023	SE	FT	...	100	ES	L
1	2023	MI	CT	...	100	US	S
2	2023	MI	CT	...	100	US	S
3	2023	SE	FT	...	100	CA	M
4	2023	SE	FT	...	100	CA	M
5	2023	SE	FT	...	0	US	L
6	2023	SE	FT	...	0	US	L
7	2023	SE	FT	...	0	CA	M
8	2023	SE	FT	...	0	CA	M
9	2023	SE	FT	...	0	US	M
10	2023	SE	FT	...	0	US	M
11	2023	SE	FT	...	100	US	M
12	2023	SE	FT	...	100	US	M
13	2023	EN	FT	...	0	US	L
14	2023	EN	FT	...	0	US	L
15	2023	SE	FT	...	0	US	M
16	2023	SE	FT	...	0	US	M
17	2023	SE	FT	...	0	US	M
18	2023	SE	FT	...	0	US	M
19	2023	MI	FT	...	100	US	M

[20 rows x 11 columns]

Hình 2: Đọc dữ liệu tùy chọn số lượng

2.1.2 Lọc dữ liệu

Với việc tập dữ liệu có rất nhiều thông tin, việc lọc dữ liệu đóng vai trò rất quan trọng, có một số thuộc tính trong trường hợp khác nhau lại không cần sử dụng đến.

Ví dụ như lọc ra những nhân viên có mức lương hơn 100000 (usd)/năm, hoặc có thể lọc ra nhân viên làm cho những công ty có quy mô là lớn và có trụ sở tại US,..

```
lọc ra các các ngành nghề có mức lương lớn hơn 100000usd
      work_year experience_level employment_type  ... remote_ratio  company_location company_size
3          2023               SE              FT  ...          100                CA            M
4          2023               SE              FT  ...          100                CA            M
5          2023               SE              FT  ...            0                US            L
6          2023               SE              FT  ...            0                US            L
7          2023               SE              FT  ...            0                CA            M
...         ...              ...             ...  ...          ...               ...          ...
3747       2021               MI              FT  ...           50                US            L
3749       2021               SE              FT  ...          100                US            L
3750       2020               SE              FT  ...          100                US            L
3751       2021               MI              FT  ...          100                US            L
3752       2020               EN              FT  ...          100                US            S

[2665 rows x 11 columns]
```

Hình 3: Kết quả lọc những nhưng viên có lương lớn hơn 100000$

Kết quả cho thấy có 2655 nhân viên có mức lương lớn hơn 100000(usd)/năm.

Dưới đây kết quả của phép lọc nhân viên làm cho công ty quy mô lớn và tại Mỹ như:Hình 4

```
Lọc ra nhân viên làm cho các công ty size L và từ US
      work_year experience_level employment_type                      job_title  ...  employee_residence  remote_ratio  company_location company_size
5          2023               SE              FT              Applied Scientist  ...                  US             0                US            L
6          2023               SE              FT              Applied Scientist  ...                  US             0                US            L
13         2023               EN              FT              Applied Scientist  ...                  US             0                US            L
14         2023               EN              FT              Applied Scientist  ...                  US             0                US            L
42         2023               EN              FT              Applied Scientist  ...                  US             0                US            L
...         ...              ...             ...                            ...  ...                 ...           ...               ...          ...
3747       2021               MI              FT  Applied Machine Learning Scientist  ...                  US            50                US            L
3749       2021               SE              FT                Data Specialist  ...                  US           100                US            L
3750       2020               SE              FT                 Data Scientist  ...                  US           100                US            L
3751       2021               MI              FT          Principal Data Scientist  ...                  US           100                US            L
3753       2020               EN              CT             Business Data Analyst  ...                  US           100                US            L

[263 rows x 11 columns]
```

Hình 4: Kết quả lọc nhiều điều kiện

Kết quả nhận được là 263 nhân viên vừa làm cho công ty có quy mô lớn và đến từ US.

Mã nguồn đọc và lọc dữ liệu bằng Python: [cautrucdulieu/doc&locdulieu.py at main · Quocdank3/cautrucdulieu · GitHub]

2.2 Gộp nhóm và thống kê cơ bản

2.2.1 Gộp nhóm

- Gộp nhóm dữ liệu theo cột "employment_type"

Đoạn code được sử dụng để đọc dữ liệu từ một tệp Excel và sau đó gộp và tính tổng các giá trị trong DataFrame theo nhóm dựa trên cột 'employment_type'. Dưới đây là mô tả cách hoạt động của đoạn code một cách rõ ràng:

1) Import thư viên pandas với viết tắt là "**pd**", thư viên này cung cấp các công cụ và chức năng để làm việc với dữ liệu dạng bảng.
2) Dùng câu lệnh "**pd.read_excel()** " để đọc dữ liệu, việc đọc dữ liệu được gán cho biến "**ch**".
3) Tạo hàm '**gop_nhom** ' nhận một DataFrame '**ch**' làm đối số.
4) Sử dụng câu lệnh '**pd.groupby()**' để gộp dữ liệu theo cột'employment_type'. Sau đó, ta sử dụng '**sum()**' để tính tổng các giá

trị trong mỗi nhóm. Kết quả là một DataFrame mới chứa kết quả gộp và tổng tương ứng.

5) Dùng '**print()**' để in kết quả đã gộp và tính tổng ra trực tiếp màn hình
6) Sử dụng '**return**' trả về kết quả là DataFrame đã gộp và tính tổng.
7) Gọi hàm '**gop_nhom(ch)**' để thực hiện gộp nhóm và tính tổng dữ liệu trong DataFrame '**ch**'.

Mã nguồn: phantichdulieu/gộp nhóm.py at main · khiemchuong/phantichdulieu · GitHub

- Gộp nhóm có cùng thuộc tính"job_title" là"Data scientist"

Đoạn code được sử dụng để đọc dữ liệu từ một tệp Excel và sau đó gộp và tính tổng các giá trị trong DataFrame dựa trên thuộc tính 'job_title' khi giá trị là 'Data Scientist'. Dưới đây là mô tả cách hoạt động của đoạn code một cách rõ ràng:

1) Import thư viên pandas với viết tắt là "**pd**", thư viên này cung cấp các công cụ và chức năng để làm việc với dữ liệu dạng bảng.
2) Dùng câu lệnh "**pd.read_excel()** " để đọc dữ liệu, việc đọc dữ liệu được gán cho biến "**ch**".
3) Tạo hàm '**gop_nhom_theo_thuoc_tinh** ' nhận một DataFrame '**ch**' làm đối số.
4) Sử dụng indexing và câu lệnh "**groupby()**" để gộp dữ liệu trong DataFrame '**ch**' theo thuộc tính 'job_title' khi giá trị là 'Data Scientist'. Sau đó, ta sử dụng '**sum()**' để tính tổng các giá trị trong mỗi nhóm. Kết quả là một DataFrame mới chứa kết quả gộp và tổng tương ứng.
5) Dùng '**print()**' để in kết quả đã gộp và tính tổng ra trực tiếp màn hình
6) Sử dụng '**return**' trả về kết quả là DataFrame đã gộp và tính tổng.
7) Gọi hàm '**gop_nhom_theo_thuoc_tinh(ch)**' để thực hiện gộp nhóm và tính tổng dữ liệu trong DataFreme '**ch**'.

Mã nguồn: phantichdulieu/gộp nhóm.py at main · khiemchuong/phantichdulieu · GitHub

2.2.2 Thống kê cơ bản

Thống kê dữ liệu bao gồm như : tập dữ liệu có bao nhiêu hàng và cột; trung bình; phân vị 25, 50, 75; phương sai và độ lệch chuẩn; giá trị lớn nhất nhỏ nhất của một thuộc tính trong tập dữ liệu; tần xuất giá trị của thuộc tính;...Để thực hiện những thống kê như trên thì có thể sử dụng Python kết hợp các thư viện như Pandas, Numpy,..

Kết quả thông kê cơ bản như: Hình 5.

```
PS E:\data\PTDL> & C:/Users/ADMIN/AppData/Local/Programs/Python/Python311/python.exe e:/data/PTDL/Do_An/thongkecoban.p
Tập dữ liệu có số hàng và cột lần lượt là (3755, 11)
Lương trung bình của mỗi năm work_year
2020     92302.631579
2021     94087.208696
2022    133338.620793
2023    149045.541176
Name: salary_in_usd, dtype: float64
Phân vị 25 của tập dữ liệu có thuộc tính về lương 95000.0
Trung vị của lương nhân viên theo USD 135000.0
Phân vị 75 của tập dữ liệu có thuộc tính về lương 175000.0
Phương sai của tổng thể về lương(USD) 3974953021.204052
Phương sai của mẫu về lương (USD) 3976011879.227814
Lương cao nhất: 450000 $
Lương thấp nhất: 5132 $
Số lượng người trong mỗi khoảng lương
 120000-200000    1690
80000-120000      825
40000-80000       527
200000-280000     451
280000-350000      65
350000-450000      15
```

Hình 5: Kết quả thống kê cơ bản

Mã nguồn thống kê cơ bản: [PTDL/thongkecoban.py at main · BAOTIN2004/PTDL · GitHub]

2.3 Trực quan hóa dữ liệu

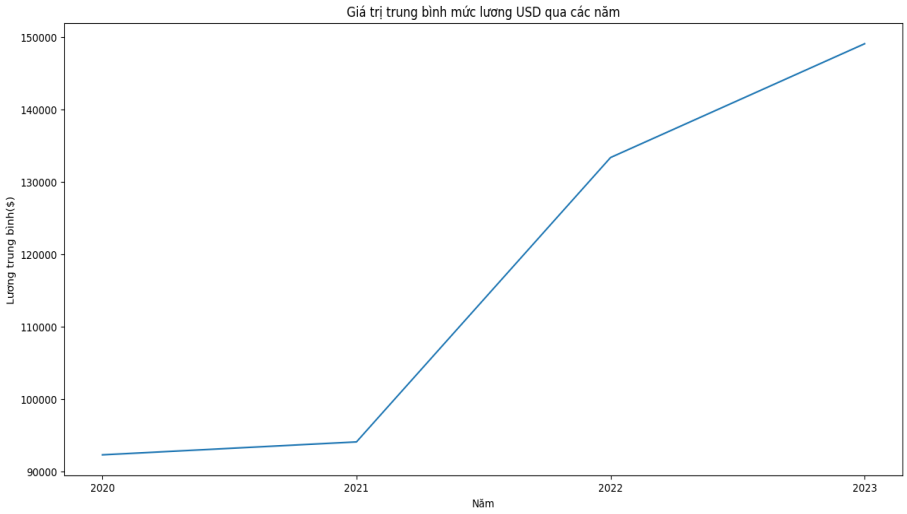
Trực quan hóa dữ liệu là kỹ thuật biểu diễn dữ liệu bằng đồ họa, có sử dụng các yếu tố trực quan như biểu đồ và đồ thị để phát hiện xu hướng, mẫu hình và các giá trị ngoại lệ, từ đó đúc kết nhanh thông tin chuyên sâu và hỗ trợ trong việc ra quyết định tức thời. Trong thế giới ngày nay, việc hiểu được khối lượng dữ liệu khổng lồ mà doanh nghiệp tạo ra mỗi ngày ngày càng quan trọng.

Để trực quan hóa dữ liệu trong Python chúng ta có thể sử dụng những thư viện Matplotlib kết hợp Numpy, Seaborn,...

2.3.1 Biểu đồ đường

Biểu đồ đường là một trong những dạng biểu đồ thông dụng, được dùng để thể hiện tiến trình phát triển, động thái phát triển của một đối tượng hay một nhóm đối tượng nào đó qua thời gian.

Dưới đây là biểu diễn lương trung bình của nhân viên qua từng năm từ 2020 đến 2023 như: Hình 6.

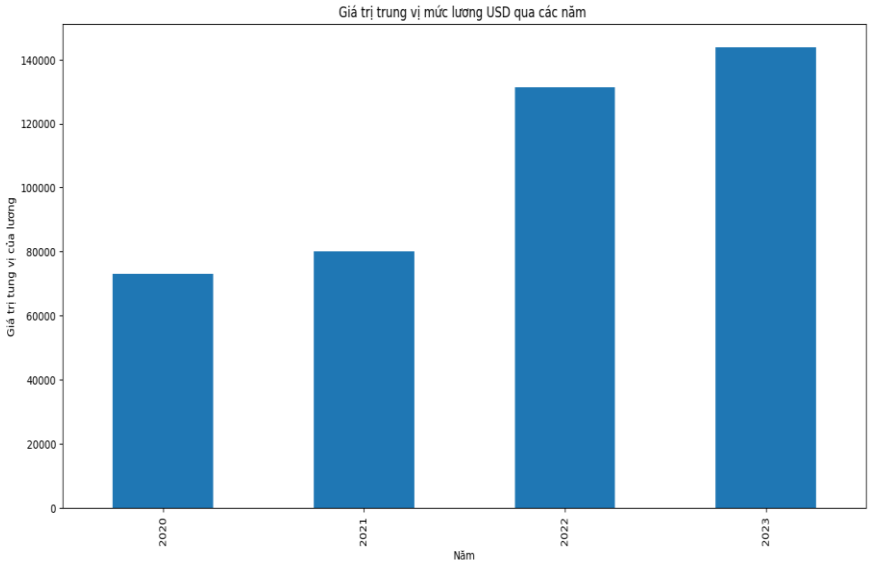


Hình 6: Biểu đồ đường

2.3.2 Biểu đồ cột

Biểu đồ cột là một dạng biểu đồ phổ biến, được dùng để thể hiện quy mô, số lượng, sản lượng, khối lượng của các đối tượng khi đề bài thường yêu cầu thể hiện tình hình phát triển, so sánh tương qua các đại lượng.

Dưới đây là kết quả vẽ biểu đồ cột giá trị trung vị của lương qua từng năm từ 2020-2023 như: Hình 7.

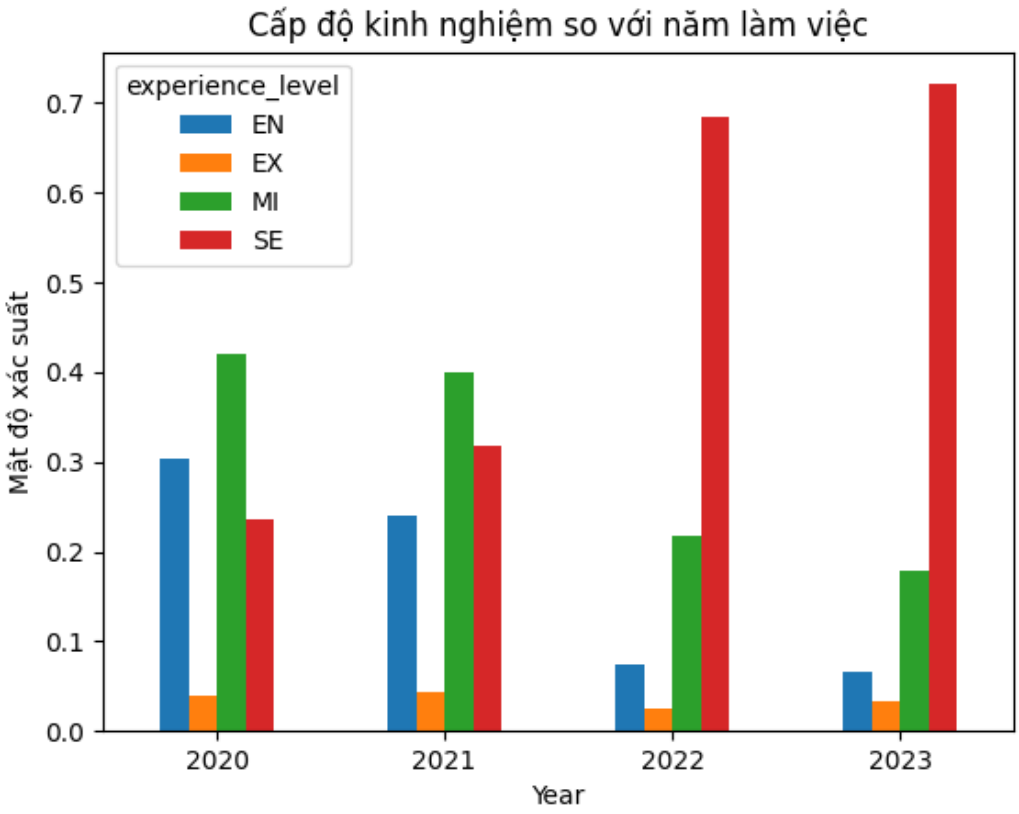


Hình 7: Biểu đồ cột

Ngoài biểu đồ cột đơn thì còn có biểu đồ cột ghép, là một loại biểu đồ sử dụng các cột để hiển thị dữ liệu trong một tập hợp các nhóm. Nó cho phép so sánh tổng giá trị của

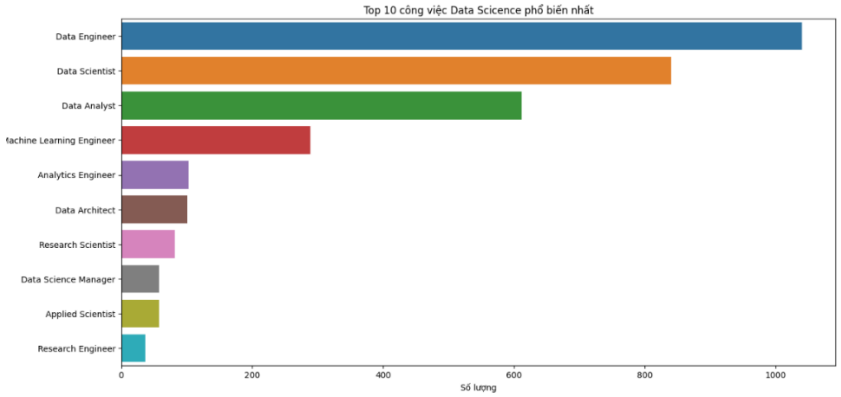
các nhóm và phần trăm mà mỗi nhóm đóng góp vào tổng số. Các cột được xếp chồng lên nhau để tạo thành một cột tổng thể.

Dưới đây là biểu đồ cột ghép thể hiện mật độ từng mức kinh nghiệm làm việc của mỗi năm như



Hình 8:Biểu đồ cột ghép

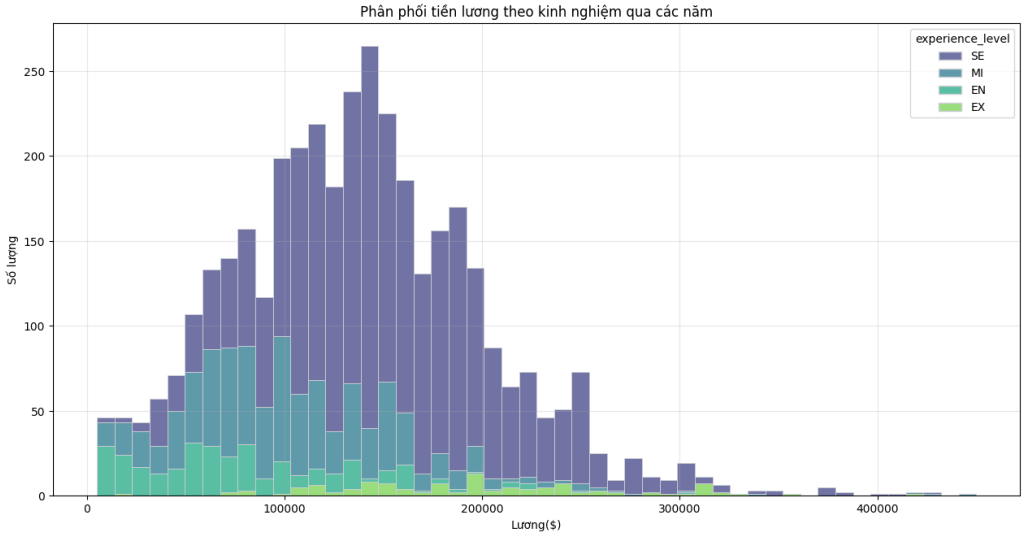
Một hình thức khác của biểu đồ cột mà được sử dụng nhiều trong việc phân tích dữ liệu, thống kê dữ liệu đó là biểu đồ thanh, dưới đây là thống kê được top 10 công việc Data Scicence như: Hình 9.



Hình 9: Biểu đồ thanh

Như vậy chúng ta có thể dễ dàng nhận định được những ngành nghề thịnh hành của công việc liên quan đến dữ liệu.

Mở rộng của biểu đồ cột là biểu đồ cột chồng, dưới đây là biểu đồ xếp chồng theo mức lương và mức độ kinh nghiệm để trực quan hóa phân phối tần suất của các mức lương dựa trên từng nhóm kinh nghiệm khác nhau như: Hình 10.

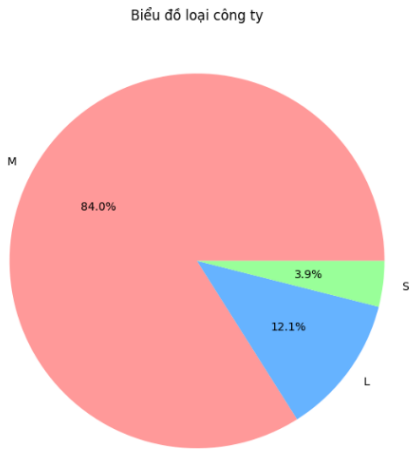


Hình 10: Biểu đồ cột chồng

2.3.3 Biểu đồ tròn

Biểu đồ tròn là dạng biểu đồ thường được dùng để vẽ các biểu đồ liên quan đến cơ cấu, tỷ lệ các thành phần trong một tổng thể chung hoặc cũng có thể vẽ biểu đồ tròn khi tỷ lệ % trong bảng số liệu cộng lại tròn 100.

Dưới đây là biểu đồ tròn thể hiện tỉ lệ quy mô công ty trong tập dữ liệu của nhóm như: Hình 11.

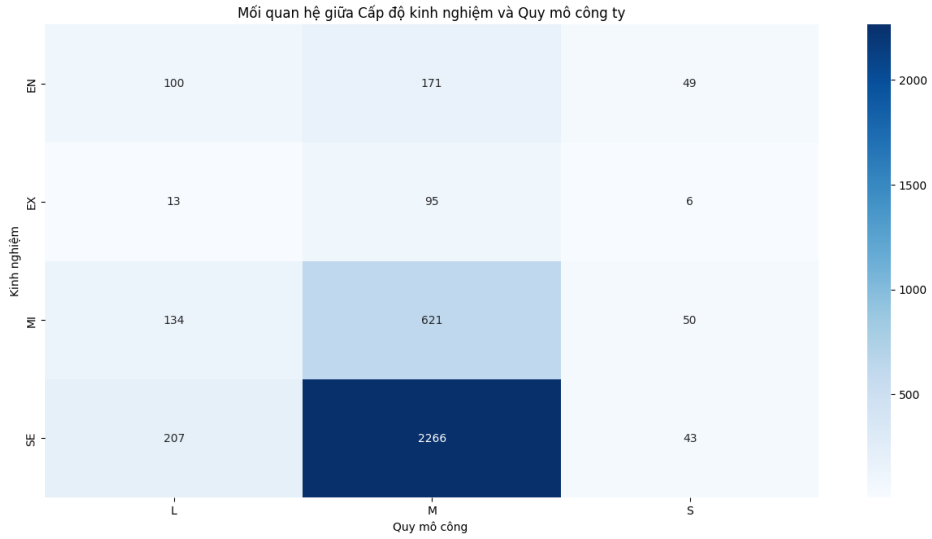


Hình 11: Biểu đồ tròn

2.3.4 Biểu đồ nhiệt

Biểu đồ nhiệt, còn được gọi là heatmap, là một biểu đồ mà sự biến thiên của các giá trị dữ liệu được biểu diễn thông qua màu sắc trên một lưới 2 chiều. Biểu đồ này thường được sử dụng để hiển thị mối quan hệ và sự tương đồng giữa các giá trị dữ liệu.

Dưới đây là biểu đồ nhiệt cho biết thông tin số lượng nhân viên từng mức độ kinh nghiệm theo từng năm như: Hình 12.



Hình 12: Biểu đồ nhiệt

Mã nguồn trực quan hóa dữ liệu của nhóm: [cautrucdulieu/vedothidulieu.py at main · Quocdank3/cautrucdulieu · GitHub]

PHẦN 3: BÀI TOÁN TƯƠNG QUAN, ƯỚC LƯỢNG, KIỂM ĐỊNH

3.1 Tương quan

Đề bài: Tính hệ số tương quan giữa lương ngành "Data Analytics Manager" với kinh nghiệm làm việc. Nhận xét về sự tương quan đó?

Trong 4 mức kinh nghiệm trong dữ liệu không phải là dạng số, vì vậy mỗi mức độ sẽ được quy định trong đoạn [1,4] để phục vụ việc tính hệ số tương quan

Dưới đây là kết quả bài toán tương quan như: Hình 13.

```
C:\Users\ADMIN\AppData\Local\Programs\Python\Python311\python.exe E:\data\PTDL\Do_An\tuongquan.py
Hệ số tương quan giữa ngành 'Data Analytics Manager' với kinh nghiệm làm việc:  0.19922425439207378

Process finished with exit code 0
```

Hình 13: Kết quả bài toán tương quan

Kết quả hệ số tương quan 0.199, dựa trên hệ số tương quan này, chúng ta không thể kết luận rằng kinh nghiệm làm việc có một tác động lớn đến lương ngành "Data Analytics Manager". Bởi vì còn phụ thuộc vào hình thức làm việc(full time,...) và tỉ lệ làm việc từ xa.

Mã nguồn bài toán tương quan: PTDL/tuongquan.py at main · BAOTIN2004/PTDL · GitHub

3.2 Ước lượng

Đề bài: Ước lượng khoảng mức lương trung bình tính bằng USD cho loại hình công việc full time, với độ tin cậy là 95% ?

Dưới đây là kết quả bài toán ước lượng sử dụng Python để giải quyết như

```
PS E:\data\PTDL> & C:/Users/ADMIN/AppData/Local/Programs/Python/Python311/python.exe "e:/data/PTDL/chương 5/1.py"
Nhập giá trị tin cậy (từ 0 đến 1): 0.95
Khoảng tin cậy của giá trị trung bình cho hình thức làm việc toàn thời gian:
(136079.41816069977, 140548.98097862245)
```

Hình 14: Kết quả bài toán ước lượng

Mã nguồn bài toán ước lượng: cautrucdulieu/uocuong_salaryusd_fulltime.py at main · Quocdank3/cautrucdulieu · GitHub

3.3 Kiểm định

Đề bài: Có phát biểu nhận định rằng có sự khác nhau giữa lương trung bình của ngành " Data Analyst" với ngành "Data Engineer". Hãy kiểm định nhận định đó mới mức ý nghĩa 5% ?

Dưới đây là kết quả bài toán kiểm định sử dụng Python để giải quuyết:

```
PS E:\data\PTDL> & C:/Users/ADMIN/AppData/Local/Programs/Python/Python311/python.exe e:/data/PTDL/Do_An/kiemdinh.py
Có sự khác biệt ý nghĩa về mặt thống kê lương trung bình giữa nhóm 'Data Analyst' và nhóm 'Data Engineer'.
PS E:\data\PTDL>
```

Hình 15: Kết quả bài toán kiểm định

Mã nguồn bài toán kiểm định: PTDL/kiemdinh.py at main · BAOTIN2004/PTDL · GitHub