

BỘ CÔNG THƯƠNG
TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI



BÁO CÁO BÀI TẬP LỚN

MÔN HỌC: PHÂN TÍCH DỮ LIỆU

**ĐỀ TÀI: PHÂN TÍCH DỮ LIỆU VÀ DỰ BÁO MỨC LƯƠNG
CỦA NGÀNH KHOA HỌC DỮ LIỆU BẰNG PHƯƠNG PHÁP
HỒI QUY**

LỚP: CH HTTT K13.1

Nhóm học viên thực hiện: Nhóm 5

- 1. Dương Mạnh Cường_Mã HV: 202370071**
- 2. Nguyễn Tuấn Dũng_Mã HV: 2023700093**

Hà Nội – Năm 2023

LỜI CẢM ƠN

Chúng em xin chân thành cảm ơn quý thầy, cô trường Đại Học Công Nghiệp Hà Nội đã tận tình dạy dỗ chúng em, trong đó phải kể đến quý thầy cô trong Khoa Công nghệ thông tin đã tạo điều kiện để chúng em thực hiện đề tài tiểu luận.

Đặc biệt, chúng em xin chân thành cảm ơn giảng viên hướng dẫn – TS. Nguyễn Mạnh Cường đã tận tình giúp đỡ, hỗ trợ chúng em trong quá trình thực hiện đề tài. Cung cấp cho chúng em những kiến thức quý báu cũng như những lời khuyên hữu ích. Tạo động lực cho chúng em hoàn thành tốt nhiệm vụ của mình. Bên cạnh đó, chúng em cũng xin cảm ơn các bạn học viên trong Khoa Công nghệ thông tin đã đóng góp ý kiến giúp chúng em thực hiện đề tài đạt hiệu quả hơn.

Bài tiểu luận này đã giúp chúng em rèn luyện kỹ năng tư duy phân tích, xử lý dữ liệu và trình bày thông tin một cách có logic và rõ ràng. Chúng em hi vọng rằng những kiến thức và kinh nghiệm thu thập từ đề tài này sẽ tiếp tục hỗ trợ chúng em trong tương lai, không chỉ trong học tập mà còn trong sự nghiệp và cuộc sống.

Một lần nữa, chúng em xin chân thành cảm ơn sự hướng dẫn và định hướng của quý thầy cô và các bạn học viên khoa Công nghệ thông tin. Chúng em rất mong nhận được những ý kiến đóng góp để đề tài được hoàn thiện hơn.

Nhóm học viên thực hiện

Dương Mạnh Cường
Nguyễn Tuấn Dũng

MỤC LỤC

LỜI CẢM ƠN.....	i
MỤC LỤC.....	ii
DANH MỤC HÌNH ẢNH.....	iv
DANH MỤC BẢNG BIỂU.....	v
DANH MỤC TỪ VIẾT TẮT	vi
LỜI NÓI ĐẦU	1
CHƯƠNG 1. TỔNG QUAN VỀ ĐỀ TÀI	2
1.1. Tổng quan về phân tích dữ liệu	2
1.1.1. Phân tích dữ liệu là gì.....	2
1.1.2. Quy trình phân tích dữ liệu.....	2
1.2. Tổng quan về bài toán dự báo.....	3
1.2.1. Lịch sử về bài toán dự báo	3
1.2.2. Tình hình nghiên cứu trong nước	4
1.2.3. Tình hình nghiên cứu ở nước ngoài	5
1.3. Bài toán phân tích dữ liệu và dự báo mức lương của ngành Khoa học dữ liệu bằng phương pháp hồi quy.....	5
1.4. Kết luận chương 1.....	7
CHƯƠNG 2. CÁC PHƯƠNG PHÁP KỸ THUẬT.....	8
2.1. Phương pháp phân tích mô tả	8
2.1.1. Phân tích mô tả	8
2.1.2. Phương pháp phân tích trên từng biến.....	9
2.1.3. Phương pháp phân tích trên nhiều biến.....	10
2.2. Phương pháp phân tích hồi quy	11

2.2.1. Tổng quan về phân tích hồi quy	11
2.2.2. Các phương pháp phân tích hồi quy	11
2.2.3. Lựa chọn phương pháp	12
2.3. Công cụ phục vụ thực hiện bài toán	12
2.3.1. Python [3]	12
2.3.2. R [4]	13
2.3.3. Lựa chọn công cụ	14
2.4. Kết luận chương 2	15
CHƯƠNG 3. THỰC NGHIỆM VÀ ĐÁNH GIÁ	16
3.1. Dữ liệu thực nghiệm	16
3.2. Quy trình thực nghiệm	17
3.2.1. Đặt mục tiêu	18
3.2.2. Tiền xử lý dữ liệu	18
3.2.3. Phân tích mô tả	21
3.2.4. Phân tích hồi quy	26
3.3. Đánh giá & Đề xuất	27
3.4. Kết luận chương 3	28
KẾT LUẬN	29
TÀI LIỆU THAM KHẢO	30

DANH MỤC HÌNH ẢNH

Hình 1.1 Quy trình phân tích dữ liệu	2
Hình 2.1 Ngôn ngữ lập trình Python.....	12
Hình 2.2 Ngôn ngữ lập trình R.....	13
Hình 3.1. Quy trình thực nghiệm đề tài phân tích dữ liệu	17
Hình 3.2. Thông tin tóm lược dữ liệu của cột dữ liệu dạng số	18
Hình 3.3. Thông tin tỷ lệ thiếu của data và tổng số data trùng	19
Hình 3.4. Quy tắc chuyển đổi cách thức biểu diễn	20
Hình 3.5. Biểu đồ hình tròn phân bố lĩnh vực làm việc.....	21
Hình 3.6. Biểu đồ phân phối lương tính theo USD.....	22
Hình 3.7. Biểu đồ Boxplot của lương tính theo USD và chế độ làm việc.....	23
Hình 3.8. Biểu đồ Heatmap của trung bình lương (USD) qua từng năm với mỗi mảng làm việc	24
Hình 3.9. Biểu đồ phân tán mức lương ảnh hưởng bởi vị trí công ty và nơi ở nhân viên	25
Hình 3.10. Biểu đồ địa lý thể hiện mức lương trung bình theo vị trí công ty.	26
Hình 3.11. Điểm R-square trên Terminal	27

DANH MỤC BẢNG BIỂU

Bảng 2.1 So sánh ưu nhược điểm để lựa chọn công cụ	14
Bảng 3.1. 10 dòng đầu của bộ dữ liệu gốc	16
Bảng 3.2. 10 dòng đầu của bộ dữ liệu sau khi chuyển đổi	21

DANH MỤC CÁC KÝ HIỆU, TỪ VIẾT TẮT

Ký hiệu, từ viết tắt	Viết đầy đủ
GDP	Gross domestic product
IEEE	the Institute of Electrical and Electronics Engineers
KDE	Kernel Density Estimation
ISO	International Organization for Standardization

LỜI NÓI ĐẦU

Trong thời đại số hóa và phát triển công nghệ như hiện nay, dữ liệu đã trở thành một tài nguyên vô cùng quý báu và quyết định cho sự phát triển của nhiều lĩnh vực. Kết hợp với sự tiến bộ trong lĩnh vực phân tích dữ liệu, khả năng khai thác thông tin từ dữ liệu ngày càng mạnh mẽ, đã mở ra những cơ hội mới cho việc hiểu rõ hơn về nhiều khía cạnh của xã hội và kinh tế. Trong bối cảnh này, việc áp dụng phân tích dữ liệu để dự báo mức lương trong các ngành nghề đang nhận được sự quan tâm đặc biệt.

Lĩnh vực Khoa học dữ liệu (Data Science) đang trở thành một trong những ngành hấp dẫn và tiềm năng với vai trò quan trọng trong việc chuyển đổi dữ liệu thành thông tin có giá trị. Trong quá trình tạo ra thông tin từ dữ liệu, việc hiểu rõ yếu tố ảnh hưởng đến mức lương trong ngành Khoa học dữ liệu đóng vai trò quan trọng để hỗ trợ quyết định tuyển dụng, phát triển sự nghiệp và định hình chiến lược nhân sự.

Trong bối cảnh này, đề tài "Phân tích dữ liệu và dự báo mức lương của ngành Khoa học dữ liệu bằng phương pháp hồi quy" được xem là hữu ích trong việc áp dụng phân tích dữ liệu để định hình tương lai cho ngành này. Bằng việc xây dựng mô hình hồi quy và phân tích các yếu tố ảnh hưởng, đề tài này hứa hẹn sẽ cung cấp cái nhìn sâu hơn về tầm quan trọng của các yếu tố như kinh nghiệm, trình độ học vấn, vị trí công việc và vùng địa lý đối với mức lương của những người làm trong lĩnh vực Khoa học dữ liệu.

Qua việc tiến hành phân tích và dự báo, đề tài này mong muốn góp phần đưa ra thông tin hữu ích cho các cá nhân quan tâm đến lĩnh vực Khoa học dữ liệu, từ các nhà quản lý tuyển dụng đến những người đang nắm giữ vai trò quan trọng trong việc quản lý nhân sự và phát triển nguồn nhân lực.

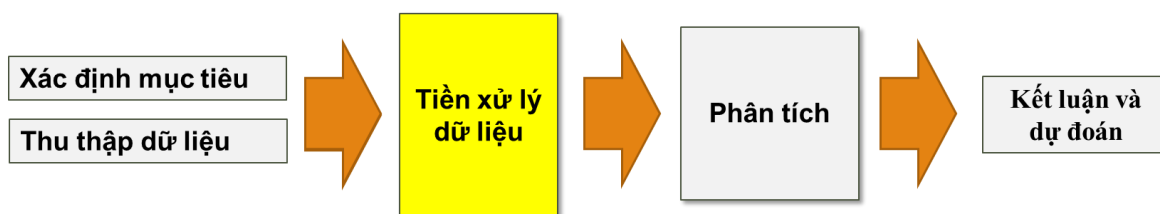
CHƯƠNG 1. TỔNG QUAN VỀ ĐỀ TÀI

1.1. Tổng quan về phân tích dữ liệu

1.1.1. Phân tích dữ liệu là gì

Phân tích dữ liệu là quá trình kiểm tra, làm sạch, chuyển đổi và mô hình hóa dữ liệu với mục tiêu khám phá thông tin hữu ích, đưa ra kết luận và hỗ trợ việc ra quyết định.

1.1.2. Quy trình phân tích dữ liệu



Hình 1.1 Quy trình phân tích dữ liệu

Quy trình phân tích dữ liệu thường bao gồm các bước chính:

- **Xác định mục tiêu và thu thập dữ liệu:**
 - + **Xác định mục tiêu:** là những kết quả cụ thể mà ta muốn đạt được thông qua việc xử lý và phân tích dữ liệu. Mục tiêu này xác định hướng đi và phạm vi của quá trình phân tích, giúp ta tập trung vào việc thu thập thông tin quan trọng và thực hiện các phân để đáp ứng các yêu cầu hoặc nhu cầu cụ thể.
 - + **Thu thập dữ liệu:** là thu thập dữ liệu từ các nguồn khác nhau như cơ sở dữ liệu, tệp tin, trang web, thiết bị cảm biến, và nhiều nguồn khác. Dữ liệu có thể là số liệu, văn bản, hình ảnh, hoặc âm thanh.
- **Tiền xử lý dữ liệu:** Dữ liệu thường không hoàn hảo và có thể chứa nhiều, dữ liệu bị thiếu, hoặc không chính xác. Tiền xử lý dữ liệu bao gồm việc tóm lược dữ liệu, làm sạch dữ liệu, tích hợp dữ liệu, chuyển đổi dữ liệu, rút gọn dữ liệu và rời rạc hóa dữ liệu để chuẩn bị cho bước phân tích.
- **Phân tích dữ liệu:** Bước quan trọng này dựa vào kiến thức và kỹ thuật phân tích để tìm ra mối liên hệ và thông tin ẩn sau dữ liệu. Phân tích dữ liệu có thể sử dụng các phương pháp phân tích mô tả, phân tích hồi quy,

phân tích sự khác biệt, thống kê, machine learning, data mining, và nhiều kỹ thuật khác.

- **Kết luận và dự đoán:** Dựa trên phân tích và thông tin từ dữ liệu, chúng ta có thể rút ra kết luận, hiểu rõ hơn về tình hình, và thậm chí đưa ra dự đoán cho tương lai.

1.2. Tổng quan về bài toán dự báo

1.2.1. Lịch sử về bài toán dự báo

Bài toán dự báo có một lịch sử lâu đời và đã phát triển qua nhiều giai đoạn. Dưới đây là một cái nhìn tổng quan về lịch sử hình thành của bài toán dự báo:

- **Thời kỳ tiền Công nghiệp (Trước thế kỷ 18):** Trong giai đoạn này, con người thường dự báo dựa trên kinh nghiệm và tri thức truyền đạt qua thế hệ. Dự báo chủ yếu dựa trên sự quan sát của thiên văn học, thời tiết, và các hiện tượng tự nhiên.

- **Cách mạng Công nghiệp và thống kê (Thế kỷ 18 - 19):** Trong thời kỳ này, việc sử dụng số liệu và thống kê để dự báo đã trở nên phổ biến hơn. Những ý tưởng về xác suất và phân phối bắt đầu được áp dụng vào việc dự báo.

- **Thế kỷ 20 và Kỹ thuật số hoá:** Sự phát triển của máy tính và kỹ thuật số hoá đã mở ra những cơ hội mới trong việc dự báo. Các phương pháp thống kê, mô hình hóa toán học, và kỹ thuật machine learning bắt đầu được sử dụng rộng rãi để dự báo trong nhiều lĩnh vực.

- **Thống kê Bayes và Kỹ thuật Machine learning (Thế kỷ 20 - 21):** Thống kê Bayes và các kỹ thuật machine learning như học máy, học sâu, và học tăng cường đã thúc đẩy khả năng dự báo thông qua việc xử lý dữ liệu phức tạp và tìm ra các mẫu ẩn.

- **Dự báo trong thời đại số hóa (Hiện nay):** Với sự gia tăng mạnh mẽ về khả năng tính toán, khối lượng dữ liệu khổng lồ, và sự phát triển của trí tuệ nhân tạo, bài toán dự báo đang trở nên càng quan trọng và phức tạp hơn. Các

công nghệ mới như big data analytics, deep learning, và dự báo dựa trên mạng xã hội đang mở ra nhiều cơ hội và thách thức mới trong lĩnh vực này.

Trong suốt quá trình phát triển, bài toán dự báo đã chuyển từ việc dự đoán dựa trên sự quan sát đơn thuần đến việc sử dụng các phương pháp phức tạp để xác định mối quan hệ phức hợp và xu hướng từ dữ liệu. Lịch sử hình thành này thể hiện sự tiến bộ và tầm quan trọng của bài toán dự báo trong việc hỗ trợ quyết định và phát triển trong nhiều lĩnh vực.

Bài toán dự báo là một trong những thách thức quan trọng trong lĩnh vực phân tích dữ liệu, nơi chúng ta cố gắng dự đoán giá trị của một biến mục tiêu trong tương lai dựa trên dữ liệu lịch sử và các yếu tố ảnh hưởng. Mục tiêu chính của bài toán dự báo là xây dựng một mô hình có khả năng hiểu và ứng dụng các mẫu, xu hướng và quy luật từ dữ liệu để thực hiện việc dự đoán một cách chính xác và đáng tin cậy.

1.2.2. Tình hình nghiên cứu trong nước

Bài toán dự báo có sự ảnh hưởng to lớn tại cả Việt Nam. Dự báo giúp cải thiện quản lý, định hình chiến lược, và tối ưu hóa tài nguyên trong nhiều lĩnh vực. Có một số điểm đáng chú ý về tình hình phân tích dữ liệu tại Việt Nam:

- **Phát triển đang ở giai đoạn đầu:** Trong một số lĩnh vực, bài toán dự báo tại Việt Nam đang ở giai đoạn đầu của sự phát triển. Việc áp dụng các phương pháp phân tích dữ liệu và dự báo mới còn đang được tìm hiểu và thí nghiệm.
- **Ứng dụng trong nông nghiệp và kinh tế:** Tại Việt Nam, dự báo có ứng dụng quan trọng trong nông nghiệp, nhằm dự đoán thời tiết, mùa màng, và nhu cầu năng lượng. Nó cũng được áp dụng trong kinh tế, dự báo tăng trưởng GDP, lạm phát, và tỷ giá.
- **Thách thức từ dữ liệu:** Một thách thức cho việc dự báo tại Việt Nam là khả năng thu thập và quản lý dữ liệu chất lượng. Dữ liệu thường không đầy đủ và có thể gặp vấn đề về tính nhất quán và độ tin cậy.

1.2.3. Tình hình nghiên cứu ở nước ngoài

Trong lĩnh vực nghiên cứu bài toán dự báo đã có một số công trình nghiên cứu ngoài nước có liên quan đến đề tài tiểu luận, ví dụ như:

“Solar Forecast Reconciliation and Effects of Improved Base Forecasts” được đăng trên IEEE Xplore, tác giả: Gokhan Mert Yagli, Dazhi Yang, Dipti Srinivasan, Monika. Đề tài nghiên cứu này trình bày về dự báo sản lượng điện mặt trời đóng vai trò quan trọng trong vận hành hệ thống điện. Dự báo được yêu cầu trên các quy mô địa lý và thời gian khác nhau, có thể được mô hình hóa dưới dạng phân cấp. [1]

Từ đó ta thấy tại nước ngoài có những sự khác biệt về bài toán dự báo:

- **Phát triển mạnh:** Tại các quốc gia phát triển, bài toán dự báo đã được phát triển mạnh và có sự ứng dụng rộng rãi trong nhiều lĩnh vực như tài chính, thương mại điện tử, y tế, và năng lượng.
- **Sự kết hợp của công nghệ mới:** Các quốc gia nước ngoài thường kết hợp sự phát triển của công nghệ mới như trí tuệ nhân tạo, học máy, và big data analytics để cải thiện hiệu suất của bài toán dự báo.
- **Tổng hợp dữ liệu:** Một ưu điểm của các quốc gia phát triển là có khả năng tổng hợp dữ liệu từ nhiều nguồn khác nhau, tạo nền tảng cho việc dự báo chính xác hơn và đa dạng hơn.

1.3. Bài toán phân tích dữ liệu và dự báo mức lương của ngành Khoa học dữ liệu bằng phương pháp hồi quy

Bài toán "Phân tích dữ liệu và dự báo mức lương của ngành Khoa học dữ liệu bằng phương pháp hồi quy" là một đề tài trong lĩnh vực phân tích dữ liệu, tập trung vào việc hiểu và dự đoán mức lương của ngành Khoa học Dữ liệu. Bài toán này đặt ra mục tiêu xác định các yếu tố ảnh hưởng đến mức lương và sử dụng phương pháp hồi quy để xây dựng một mô hình dự báo.

- **Mục tiêu nghiên cứu:**

+ ***Phân tích yếu tố ảnh hưởng***: Hiểu rõ các yếu tố có thể ảnh hưởng đến mức lương của ngành Khoa học Dữ liệu. Các yếu tố này có thể là học vấn, kinh nghiệm làm việc, vị trí công việc, vùng địa lý, và các yếu tố khác.

+ ***Xây dựng mô hình hồi quy***: Sử dụng phương pháp hồi quy để xây dựng mô hình dự báo mức lương dựa trên các yếu tố ảnh hưởng đã được xác định. Mô hình hồi quy sẽ cố gắng tìm ra mối quan hệ giữa các biến độc lập và biến phụ thuộc (mức lương).

+ ***Dự đoán mức lương***: Dựa trên mô hình hồi quy đã xây dựng, mục tiêu là dự đoán mức lương cho những cá nhân có các thông tin liên quan đã được cung cấp.

- **Ý nghĩa khoa học và thực tiễn:**

+ ***Khoa học dữ liệu***: Đề tài này đóng góp vào lĩnh vực Khoa học Dữ liệu bằng cách áp dụng các kỹ thuật phân tích dữ liệu và hồi quy để khám phá mối liên hệ giữa yếu tố ảnh hưởng và mức lương, từ đó cung cấp thông tin giá trị về ngành và thị trường lao động.

+ ***Quản lý nhân sự***: Kết quả của nghiên cứu có thể giúp các công ty và tổ chức trong ngành Khoa học Dữ liệu hiểu rõ hơn về các yếu tố ảnh hưởng đến mức lương của nhân viên. Điều này có thể hỗ trợ trong việc đưa ra quyết định về chiến lược tuyển dụng, phát triển nhân viên, và quản lý tài nguyên.

+ ***Tư duy phân tích***: Việc thực hiện phân tích dữ liệu và xây dựng mô hình hồi quy trong ngữ cảnh của bài toán này cũng giúp phát triển kỹ năng tư duy phân tích, sáng tạo, và khả năng áp dụng các phương pháp phân tích vào các vấn đề thực tế.

Như vậy, bài toán này không chỉ có ý nghĩa đối với lĩnh vực Khoa học Dữ liệu mà còn mang lại những kiến thức hữu ích cho nhiều khía cạnh khác trong xã hội và kinh tế.

1.4. Kết luận chương 1

Chương 1 đã trình bày tổng quan về đề tài, trình bày tổng quan về phân tích dữ liệu và bài toán dự báo, đồng thời nêu ra tình hình nghiên cứu bài toán dự báo trong nước và ở nước ngoài. Ngoài ra còn mô tả bài toán phân tích dữ liệu và dự báo mức lương của ngành Khoa học dữ liệu bằng phương pháp hồi quy.

CHƯƠNG 2. CÁC PHƯƠNG PHÁP KỸ THUẬT

2.1. Phương pháp phân tích mô tả

2.1.1. Phân tích mô tả

Phân tích mô tả là một phương pháp trong lĩnh vực thống kê và phân tích dữ liệu, nhằm mô tả và tóm tắt các đặc điểm chính của một tập dữ liệu một cách dễ hiểu và ngắn gọn. Mục tiêu của phân tích mô tả là giúp hiểu sâu hơn về dữ liệu mà chúng ta đang làm việc, nhận ra các đặc trưng quan trọng, và cung cấp một cái nhìn tổng quan về phân phối và biến đổi của dữ liệu.

Phân tích mô tả thường bao gồm các khía cạnh sau:

- **Thống kê tóm tắt:** Đây là các số liệu thống kê cơ bản như trung bình, trung vị, độ lệch chuẩn, và phân vị. Các số liệu này giúp ta hiểu về trung tâm và phân tán của dữ liệu.
- **Biểu đồ:** Biểu đồ thường được sử dụng để biểu diễn dữ liệu một cách trực quan. Các biểu đồ như biểu đồ cột, biểu đồ đường, biểu đồ hình tròn, và biểu đồ hộp giúp ta thấy được sự phân bố và xu hướng của dữ liệu.
- **Phân phối dữ liệu:** Phân tích phân phối dữ liệu giúp ta hiểu về tỷ lệ xuất hiện của các giá trị khác nhau trong tập dữ liệu. Điều này có thể làm bằng cách tạo biểu đồ phân phối tần số hoặc xây dựng biểu đồ kernel density.
- **Kiểm tra sự tương quan:** Phân tích mô tả cũng có thể liên quan đến việc kiểm tra sự tương quan giữa các biến. Điều này có thể thực hiện bằng cách sử dụng biểu đồ tương quan hoặc tính toán hệ số tương quan Pearson.
- **Xác định điểm ngoại lệ:** Phân tích mô tả cũng giúp xác định các điểm dữ liệu ngoại lệ, tức là những giá trị rất khác biệt so với phần còn lại của dữ liệu.
- **Tổng kết và nhận xét:** Cuối cùng, phân tích mô tả thường đi kèm với việc tổng kết và nhận xét về các đặc điểm quan trọng của dữ liệu, những mẫu thú vị, và những điểm mạnh và điểm yếu của tập dữ liệu.

Phân tích mô tả giúp xây dựng một cái nhìn sâu hơn về tập dữ liệu ban đầu và tạo nền tảng cho các phân tích tiếp theo như dự báo, phân tích hồi quy, hay machine learning.

2.1.2. Phương pháp phân tích trên từng biến

Khi thực hiện phân tích trên một biến (hoặc một thuộc tính), mục tiêu chính là hiểu rõ các đặc điểm cơ bản của biến đó. Điều này thường bao gồm xác định và xử lý các giá trị ngoại lai hoặc bất thường (Outliers). Đây là các giá trị dữ liệu mà rất khác biệt so với phần lớn các giá trị khác trong tập dữ liệu. Các giá trị ngoại lai có thể xuất hiện do lỗi nhập liệu, lỗi đo lường, hoặc đơn giản là do các sự kiện hiếm gặp.

Việc xác định các Outliers có vai trò quan trọng và là mắt xích liên kết giữa phân tích mô tả và phân tích hồi quy, bởi vì ta có thể tiến hành làm sạch những giá trị này tại công đoạn tiền xử lý dữ liệu của phân tích hồi quy. Cụ thể với từng loại dữ liệu khác nhau, ta sẽ phân tích như sau:

- **Dữ liệu số:**

- + **Biểu đồ Histogram:** Biểu đồ hiển thị tần suất xuất hiện của các khoảng giá trị dữ liệu.

- + **Các đại lượng thống kê:** Bao gồm mean (trung bình), stdev (độ lệch chuẩn), median (trung vị), quartile (phân vị) ... Các giá trị này giúp mô tả trung bình, phương sai và phân phối của dữ liệu.

- + **Biểu đồ Box & Whisker (Boxplot):** Biểu đồ hiển thị tổng quan giá trị đó bao gồm các giá trị đại lượng thống kê đã tính được.

- **Dữ liệu phi số:**

- + **Bảng tần suất (Frequency table):** Biểu đồ liệt kê các giá trị khác nhau của biến và số lần xuất hiện của mỗi giá trị.

- + **Biểu đồ cột (Bar chart):** Biểu đồ thể hiện tần suất của từng giá trị dữ liệu dưới dạng các cột đứng.

- + **Biểu đồ hình tròn hoặc donut (Pie chart, Donut chart):** Biểu đồ thể hiện phần trăm tần suất của từng giá trị trong tổng số.

2.1.3. Phương pháp phân tích trên nhiều biến

Phân tích trên nhiều biến hướng tới việc hiểu mối quan hệ và tương tác giữa các biến trong tập dữ liệu. Điều này có thể giúp bạn phát hiện ra các mẫu, xu hướng hoặc tương quan có thể tồn tại giữa chúng.

Các mối liên hệ giữa các biến (Interrelationships) có thể là nhiều dạng khác nhau: Mối tương quan tuyến tính, tương quan không tuyến tính, tương quan ngược... Với mỗi mối liên hệ, ta có thể phân tích và tìm ra được cách các biến tương tác và ảnh hưởng lẫn nhau.

Việc phân tích trên nhiều biến cũng có mối liên hệ mật thiết đến phân tích hồi quy khi giúp ta xác định được các giá trị ngoại lai của dữ liệu. Do là phân tích nhiều biến, vậy nên sẽ có 3 kiểu dữ liệu phân tích khác nhau: Số, phi số và hỗn hợp (cả số và phi số):

- **Dữ liệu số:**

- + ***Scatter Plot (Biểu đồ Scatter)***: Biểu đồ thể hiện mối quan hệ giữa hai biến số. Mỗi điểm trên biểu đồ thể hiện một cặp giá trị của hai biến trên trục ngang và dọc. Biểu đồ này dùng để tìm kiếm sự tương quan giữa 2 biến số như tương quan tuyến tính hoặc không tuyến tính.

- + ***Bảng dữ liệu thống kê (Statistical Summary Table)***: Tạo bảng để liệt kê các đại lượng thống kê (mean, median, stdev...) giữa các biến số của dữ liệu.

- **Dữ liệu phi số:**

- + ***Bảng dữ liệu thống kê (Statistical Summary Table)***: Cũng là bảng dữ liệu thống kê nhưng với giá trị phi số, đó sẽ chỉ có giá trị tần suất xuất hiện (mode) của dữ liệu.

- **Dữ liệu hỗn hợp**

- + ***Bảng thống kê tổng hợp***: Đây là sự kết hợp giữa bảng dữ liệu thống kê của dữ liệu số và phi số. Sự kết hợp tổng quan này sẽ cho ta bao quát được phân bố của dữ liệu.

- + **Biểu đồ Box-and-Whisker (Boxplot):** Được sử dụng để so sánh phân phối của một dữ liệu số với tần suất của một dữ liệu phi số. Biểu đồ này sẽ cho ta mối quan hệ mật thiết về sự ảnh hưởng của các giá trị phi số lên giá trị số được phân tích.

2.2. Phương pháp phân tích hồi quy

2.2.1. Tổng quan về phân tích hồi quy

Phân tích hồi quy là một tập hợp các phương pháp thống kê được sử dụng để ước tính các mối quan hệ giữa một biến phụ thuộc và một hoặc nhiều biến độc lập. Nó có thể được sử dụng để đánh giá sức mạnh của mối quan hệ giữa các biến và để mô hình hóa mối quan hệ trong tương lai giữa chúng.

Phân tích hồi quy là một cách phân loại toán học để xác định biến nào trong số những biến đó thực sự có tác động. Nó trả lời các câu hỏi: Yếu tố nào quan trọng nhất? Cái nào có thể bỏ qua? Các yếu tố đó tương tác với nhau như thế nào? Và quan trọng nhất, chúng ta chắc chắn như thế nào về tất cả những yếu tố này? [2]

Trong phân tích hồi quy, ta cần xác định một biến phụ thuộc – yếu tố chính mà ta đang cố gắng hiểu hoặc dự đoán. Phân tích hồi quy bao gồm một số biến thể, chẳng hạn như tuyến tính, nhiều tuyến tính và phi tuyến tính. Trong đó mô hình phổ biến là tuyến tính và nhiều tuyến tính. Đối với phân tích hồi quy phi tuyến, chúng thường được sử dụng cho các tập dữ liệu phức tạp hơn trong đó các biến phụ thuộc và độc lập thể hiện mối quan hệ phi tuyến.

2.2.2. Các phương pháp phân tích hồi quy

Để phân tích hồi quy có rất nhiều phương pháp để phân tích. Dưới đây sẽ là một số phương pháp quan trọng dùng để phân tích hồi quy:

- **Hồi quy tuyến tính (Linear Regression):** Hồi quy tuyến tính dự đoán giá trị mục tiêu dựa trên biến độc lập bằng cách tìm đường thẳng "tốt nhất" vượt qua dữ liệu. Phương pháp này đơn giản và phù hợp với dữ liệu có mối quan hệ tuyến tính. Tuy nhiên, nó có thể không xử lý được dữ liệu phi tuyến và ảnh hưởng bởi nhiễu dữ liệu.

- **Hồi quy Ridge (Ridge Regression):** Hồi quy Ridge là phiên bản cải tiến của hồi quy tuyến tính bằng cách thêm hệ số điều chuẩn l2 vào hàm mất mát. Điều này giúp kiểm soát độ phức tạp của mô hình và tránh tình trạng quá khớp (overfitting). Tuy ưu điểm là giảm overfitting và xử lý đa cộng tuyến, nhưng cần lựa chọn tham số điều chuẩn chính xác.
- **Hồi quy Lasso (Lasso Regression):** Hồi quy Lasso cũng cải tiến từ hồi quy tuyến tính, nhưng thay vì l2, nó sử dụng hệ số điều chuẩn l1 để thúc đẩy một số hệ số về 0. Điều này dẫn đến lựa chọn biến tự động và giảm biến quan trọng. Lasso giải quyết vấn đề "chọn biến" nhưng cần phải có tham số điều chuẩn chính xác.

2.2.3. Lựa chọn phương pháp

Phương pháp phân tích hồi quy tuyến tính (Linear Regression) là sự kết hợp của tính đơn giản, khả năng ước lượng mối quan hệ tuyến tính, khả năng dự báo cùng với khả năng phân tích định lượng. Phương pháp này sẽ giúp ta đạt được mục tiêu nghiên cứu và trả lời những câu hỏi quan trọng. Vì vậy chúng em lựa chọn phương pháp hồi quy tuyến tính để giải quyết đề bài.

2.3. Công cụ phục vụ thực hiện bài toán

2.3.1. Python [3]



Hình 2.1 Ngôn ngữ lập trình Python

Python là một trong những ngôn ngữ lập trình phổ biến nhất hiện nay, thường được sử dụng để xây dựng trang web và phần mềm, tự động hoá các tác vụ và tiến hành phân tích dữ liệu. Với sự phát triển của khoa học dữ liệu hiện

nay, Python lại càng được ứng dụng rộng rãi hơn trong ngành Data Analyst. Với thư viện đa dạng trong các lĩnh vực như khai thác dữ liệu (Scrapy, BeautifulSoup4, ...), xử lý dữ liệu và mô hình hóa (Pandas, Scikit-learn, ...), trực quan hóa dữ liệu (Matplotlib, Plotly, ...) thì đây là một lựa chọn tuyệt vời để phân tích dữ liệu. Tuy nhiên bên cạnh những ưu điểm về thư viện cũng như cộng đồng lập trình đông đảo, Python vẫn vướng phải một số nhược điểm, đó là bị giới hạn tốc độ, mức tiêu thụ bộ nhớ cao và không phải là một ngôn ngữ được hỗ trợ nhiều cho môi trường di động.

2.3.2. R [4]



Hình 2.2 Ngôn ngữ lập trình R

Ngôn ngữ R là một ngôn ngữ lập trình và môi trường tính toán thống kê phổ biến trong lĩnh vực phân tích dữ liệu và thống kê. Nó cung cấp nền tảng mạnh mẽ cho việc thực hiện các phân tích thống kê, xử lý dữ liệu và tạo biểu đồ. R cũng là một cộng đồng mã nguồn mở lớn, điều này có nghĩa là người dùng có thể dễ dàng chia sẻ mã nguồn, gói phân tích và kiến thức với nhau. Vậy nên việc phân tích dữ liệu trên R cũng rất thuận tiện khi có đầy đủ các thư viện về phân tích dữ liệu và có khả năng tích hợp tốt với môi trường nghiên cứu khoa học. Dù vậy, R vẫn có một vài nhược điểm nhất định. Phổ biến trong số đấy là sự phức tạp của ngôn ngữ khi lập trình viên mới bắt đầu tiếp xúc và sử dụng, xử lý dữ liệu lớn không tốt so với nhiều ngôn ngữ khác và hiệu suất không phải lúc nào cũng ổn định.

2.3.3. Lựa chọn công cụ

Cả Python và R đều là hai ngôn ngữ phổ biến được sử dụng cho phân tích dữ liệu và thống kê. Việc lựa chọn sử dụng ngôn ngữ nào phụ thuộc vào nhiều yếu tố như mục tiêu, kinh nghiệm cá nhân, loại dữ liệu đang làm việc, và các thư viện hỗ trợ cần sử dụng. Sau đây là bảng so sánh để đưa ra quyết định lựa chọn công cụ phục vụ giải quyết bài toán:

Bảng 2.1 So sánh ưu nhược điểm để lựa chọn công cụ

Ngôn ngữ	Python	R
Ưu điểm	<ul style="list-style-type: none">- Đa năng: Python không chỉ giới hạn trong phân tích dữ liệu, mà còn có thể sử dụng cho nhiều mục đích khác như phát triển ứng dụng, web, automation, và machine learning.- Thư viện phong phú: Có nhiều thư viện mạnh mẽ giúp thực hiện các tác vụ phân tích và xử lý dữ liệu một cách hiệu quả.- Cộng đồng lớn: python có cộng đồng lớn giúp việc chia sẻ, học hỏi dễ dàng hơn.	<ul style="list-style-type: none">- Thống kê chuyên sâu: R được thiết kế đặc biệt cho thống kê và phân tích dữ liệu, với nhiều gói như dplyr, ggplot2, tidyr, và lubridate giúp thực hiện các tác vụ phân tích chi tiết.- Biểu đồ phức tạp: Gói ggplot2 trong R cho phép tạo ra biểu đồ phức tạp và tùy chỉnh một cách dễ dàng.
Nhược điểm	<p>Thống kê chuyên sâu: Mặc dù Python có thư viện thống kê tốt, nhưng R vẫn là lựa chọn phổ biến hơn trong các nghiên cứu thống kê và phân tích dữ liệu chuyên sâu.</p>	<ul style="list-style-type: none">- Thiếu phổ biến: R có tính chuyên môn hơn so với Python.- Sử dụng bộ nhớ: R có xu hướng sử dụng nhiều bộ nhớ hơn so với Python.- Quản lý mã nguồn: R không thể sử dụng mã nguồn

		mở rộng và phân chia mã nguồn dễ dàng như Python. Việc quản lý và tái sử dụng mã có thể trở nên khó khăn hơn khi dự án phát triển.
--	--	--

Sau khi tổng hợp các ưu, nhược điểm của cả hai ngôn ngữ, chúng em quyết định sử dụng ngôn ngữ Python với sự đa năng, cộng đồng lớn và nhiều thư viện hỗ trợ.

2.4. Kết luận chương 2

Chương 2 đã trình bày các phương pháp kỹ thuật, cụ thể là phương pháp phân tích mô tả, phương pháp phân tích hồi quy và các công cụ thực hiện bài toán. Đồng thời lựa chọn được phương pháp phân tích hồi quy tuyến tính và ngôn ngữ Python để thực hiện thực nghiệm.

CHƯƠNG 3. THỰC NGHIỆM VÀ ĐÁNH GIÁ

3.1. Dữ liệu thực nghiệm

Trong project này, bộ dữ liệu được phân tích ở đây là file dataset (.csv) chứa 6775 thông tin về công việc thuộc ngành Khoa học dữ liệu được cập nhật từ năm 2020 đến nay.[5]

Cụ thể thông tin như sau:

- Tên bộ dữ liệu: Data Science Salaries
- Nguồn: <https://ai-jobs.net/salaries/download/salaries.csv>
- Dữ liệu 10 dòng đầu của dataset:

Bảng 3.1. 10 dòng đầu của bộ dữ liệu gốc

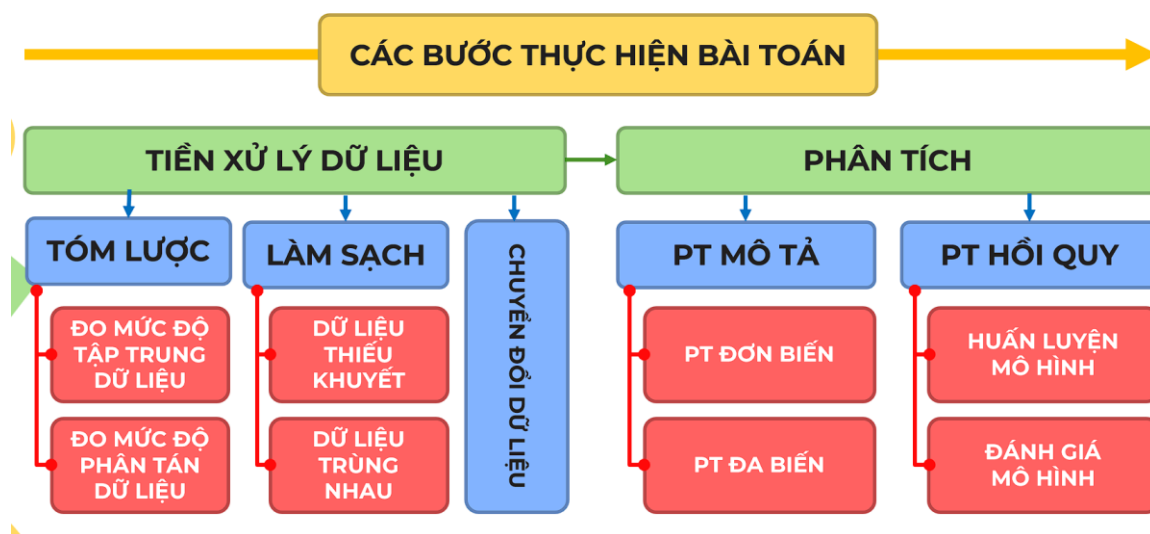
	A	B	C	D	E	F	G	H	I	J	K
1	work_year	experience_level	employment_type	job_title	salary	salary_currency	salary_in_usd	employment_type_remote	company_size	company_size	company_size
2	2023	MI	FT	AI Engineer	140000	USD	140000	US	100	US	S
3	2023	MI	FT	Data Manager	130000	USD	130000	US	0	US	M
4	2023	MI	FT	Data Manager	100000	USD	100000	US	0	US	M
5	2023	SE	FT	Machine Learning Scientist	163800	USD	163800	US	0	US	M
6	2023	SE	FT	Machine Learning Scientist	126000	USD	126000	US	0	US	M
7	2023	SE	FT	Data Scientist	225000	USD	225000	US	100	US	M
8	2023	SE	FT	Data Scientist	111967	USD	111967	US	100	US	M
9	2023	SE	FT	Data Engineer	205600	USD	205600	US	0	US	L
10	2023	SE	FT	Data Engineer	105700	USD	105700	US	0	US	L

Thông tin cụ thể các cột của dataset như sau:

- “**work_year**”: Năm làm việc của công việc đó
- “**experience_level**”: Trình độ/ Kinh nghiệm làm việc:
 - + **EN**: Entry-level / Junior (Sơ cấp)
 - + **MI**: Mid-level / Intermediate (Trung cấp)
 - + **SE**: Senior-level / Expert (Cao cấp)
 - + **EX**: Executive-level / Director (Giám đốc)
- “**employment_type**”: Chế độ làm việc:
 - + **PT**: Part-time (Bán thời gian)
 - + **FT**: Full-time (Toàn thời gian)
 - + **CT**: Contract (Hợp đồng có thời hạn)
 - + **FL**: Freelance (Làm tự do)
- “**job_title**”: Tên công việc

- **“salary”**: Tổng số lương được trả
- **“salary_currency”**: Đơn vị tiền tệ của cột **“salary”** theo chuẩn ISO 4217
- **“salary_in_usd”**: Mức lương bằng USD (Tỷ giá đã được chia theo năm tương ứng) thông qua dữ liệu thống kê từ BIS và ngân hàng trung ương.
- **“employee_residence”**: Quốc gia cư trú của nhân viên theo mã thuộc chuẩn ISO 3166
- **“remote_ratio”**: Tỷ lệ làm việc từ xa:
 - + **0**: Không làm việc từ xa (<20%)
 - + **50**: Làm việc linh hoạt
 - + **100**: Làm việc từ xa toàn thời gian (>80%)
- **“company_location”**: Quốc gia của địa điểm công ty theo mã thuộc chuẩn ISO 3166
- **“company_size”**: Số lao động bình quân làm việc:
 - + **S**: Small (Nhỏ) – ít hơn 50 người
 - + **M**: Medium (Trung bình) – từ 50 đến 250 người
 - + **L**: Large (Lớn) – nhiều hơn 250 người

3.2. Quy trình thực nghiệm



Hình 3.1. Quy trình thực nghiệm đề tài phân tích dữ liệu

3.2.1. Đặt mục tiêu

- Phân tích mô tả để thể hiện mối quan hệ giữa các giá trị của dữ liệu, từ đó đánh giá được tương quan của ngành Khoa học dữ liệu.
- Phân tích hồi quy để dự báo mức lương dựa theo mô hình hồi quy tuyến tính.

3.2.2. Tiền xử lý dữ liệu

- Tóm lược dữ liệu

Tóm lược dữ liệu trong phân tích dữ liệu là quá trình tổng hợp, trích xuất và trình bày các thông tin quan trọng và chính xác từ tập dữ liệu ban đầu. Mục tiêu của việc tóm lược dữ liệu là giúp người đọc hoặc người xem nắm bắt được những điểm quan trọng và khái quát của dữ liệu mà không cần phải đọc hoặc xem toàn bộ dữ liệu gốc. Tóm lược dữ liệu bao gồm 2 loại đo: Đo mức độ tập trung dữ liệu (mean, median, mode, ...) và Đo mức độ phân tán dữ liệu (quartile, interquartile, standard deviation, ...).

Ta sẽ tiến hành tổng hợp các thông tin về độ tập trung và phân tán của dữ liệu. Những thông số này chỉ tương thích với các cột dữ liệu dạng thông số, vậy nên sẽ chỉ có “**work_year**”, “**salary**”, “**salary_in_usd**” và “**remote_ratio**” là được phân tích. Dưới đây là kết quả tóm lược dữ liệu bao gồm các thuộc tính **count**, **mean**, **std**, **min**, **25%**, **50%**, **75%**, **max**, **mode**, **median** của các dữ liệu trên:

	work_year	salary	salary_in_usd	remote_ratio
count	6775.000000	6.775000e+03	6775.000000	6775.000000
mean	2022.659779	1.799383e+05	147883.634539	41.394834
std	0.594581	5.149935e+05	63967.970324	48.460685
min	2020.000000	1.400000e+04	15000.000000	0.000000
25%	2022.000000	1.050000e+05	104584.000000	0.000000
50%	2023.000000	1.438600e+05	141525.000000	0.000000
75%	2023.000000	1.871000e+05	185900.000000	100.000000
max	2023.000000	3.040000e+07	450000.000000	100.000000
mode	2023.000000	1.000000e+05	150000.000000	0.000000
median	2023.000000	1.438600e+05	141525.000000	0.000000

Hình 3.2. Thông tin tóm lược dữ liệu của cột dữ liệu dạng số

- Làm sạch dữ liệu:

Làm sạch dữ liệu là quá trình loại bỏ các sai sót, lỗi, nhiễu và thông tin không chính xác hoặc không cần thiết khỏi tập dữ liệu ban đầu để đảm bảo dữ liệu đáng tin cậy và phù hợp cho việc phân tích và xử lý tiếp theo. Quá trình làm sạch dữ liệu thường là một phần quan trọng trong tiền xử lý dữ liệu trước khi bắt đầu phân tích mô tả và cả phân tích hồi quy.

Một số tác vụ chính trong quá trình làm sạch dữ liệu bao gồm:

- + **Loại bỏ dữ liệu trùng lặp:** Loại bỏ các bản ghi bị trùng lặp trong tập dữ liệu để tránh ảnh hưởng đến kết quả phân tích.
- + **Xử lý dữ liệu thiếu:** Điền vào các giá trị thiếu hoặc quyết định loại bỏ chúng dựa trên ngữ cảnh và mục tiêu của phân tích.
- + **Sửa lỗi và sai sót:** Điều tra và sửa các lỗi cú pháp, sai sót chính tả hoặc sai sót logic trong dữ liệu.
- + **Chọn lọc đặc trưng:** Xác định và lựa chọn các đặc trưng quan trọng nhất để sử dụng trong phân tích hoặc mô hình hóa.

Đối với project hiện tại, sau khi khảo sát chi tiết các cột dữ liệu, việc sửa lỗi sai sót và chọn lọc đặc trưng cho dataset không quá quan trọng nên ta sẽ tiến hành loại bỏ dữ liệu trùng lặp và xử lý dữ liệu thiếu. Để làm điều này, trước hết ta cần khảo sát số data bị thiếu và trùng lặp. Kết quả khảo sát như sau:

Tỷ lệ thiếu data	
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
SỐ LƯỢNG DATA BỊ THIẾU VÀ TRÙNG LẶP: 2652	

Hình 3.3. Thông tin tỷ lệ thiếu của data và tổng số data trùng

Qua khảo sát, ta đánh giá được tài liệu không có vùng bị thiếu, tuy nhiên phần data bị trùng lặp có tổng cộng 2652 hàng dữ liệu bị trùng lặp. Ta sẽ tiến hành xóa đi phần bị thừa này để đảm bảo kết quả phân tích chính xác nhất.

- Chuyển đổi dữ liệu

Chuyển đổi dữ liệu trong phân tích dữ liệu là quá trình thay đổi cách thức biểu diễn, xử lý hoặc áp dụng các phép toán trên dữ liệu ban đầu để tạo ra dữ liệu mới có ý nghĩa hoặc thuận tiện hơn cho mục đích phân tích. Nó có vai trò quan trọng trong việc biểu diễn trực quan hơn dataset, thuận tiện hơn trong việc phân tích dữ liệu.

Trong dataset của project, ta thấy một số cột thông tin đang được ký tự viết tắt hoặc được chuẩn hóa theo một dạng khác, điều này có thể khiến người xem dữ liệu khó nắm bắt thông tin cần thiết. Việc thay đổi cách thức biểu diễn là cần thiết và ta sẽ áp dụng lên một số thông tin ở các cột **‘experience_level’**, **‘employment_type’**, **‘company_size’** và **‘remote_ratio’** cụ thể như sau:

experience_level		employment_type		company_size		remote_ratio	
SE	Senior level	FL	Freelancer	S	Small	0	On-Site
EN	Entry level	CT	Contractor	M	Medium	50	Half-Remote
EX	Executive level	FT	Full-Time	L	Large	100	Full-Remote
MI	Mid/Intermediate level	PT	Part-Time				

Hình 3.4. Quy tắc chuyển đổi cách thức biểu diễn

Các giá trị ký tự viết tắt hoặc chuẩn hóa theo dạng khác đối với các cột data trên sẽ được thay đổi thành các từ đầy đủ để dataset trực quan hơn.

Tiếp theo, để tiện lợi cho việc phân tích các lĩnh vực công việc của dataset, ta sẽ tiến hành tạo thêm một cột mới có tên **‘job_role’** dùng để phân loại các tên công việc lấy từ cột **‘job_title’** thành một số lĩnh vực cụ thể liên quan đến ngành Khoa học dữ liệu. Các lĩnh vực sẽ được phân loại bao gồm:

- + Data Engineering
- + Data Science
- + Machine Learning

- + Data Architecture
- + Management

Những công việc khác 5 lĩnh vực trên sẽ được gán giá trị ‘Other’.

Sau khi đã chuyển đổi dữ liệu, dataset mới sẽ được biểu diễn như sau:

Bảng 3.2. 10 dòng đầu của bộ dữ liệu sau khi chuyển đổi

work_year	experience_level	employment_type	job_title	salary	ry_currency	in_uae_re	remote_ratio	iny_lo	company_size	job_role	
2023	Mid/Intermediate level	Full-Time	AI Engineer	140000	USD	140000	US	Full-Remote	US	Small	Data Engineering
2023	Mid/Intermediate level	Full-Time	Data Manager	130000	USD	130000	US	On-site	US	Medium	Management
2023	Mid/Intermediate level	Full-Time	Data Manager	100000	USD	100000	US	On-site	US	Medium	Management
2023	Senior level	Full-Time	Machine Learning Scientist	163800	USD	163800	US	On-site	US	Medium	Machine Learning
2023	Senior level	Full-Time	Machine Learning Scientist	126000	USD	126000	US	On-site	US	Medium	Machine Learning
2023	Senior level	Full-Time	Data Scientist	225000	USD	225000	US	Full-Remote	US	Medium	Data Science
2023	Senior level	Full-Time	Data Scientist	111967	USD	111967	US	Full-Remote	US	Medium	Data Science
2023	Senior level	Full-Time	Data Engineer	205600	USD	205600	US	On-site	US	Large	Data Engineering
2023	Senior level	Full-Time	Data Engineer	105700	USD	105700	US	On-site	US	Large	Data Engineering
2023	Mid/Intermediate level	Full-Time	Data Manager	115500	USD	115500	US	On-site	US	Medium	Management

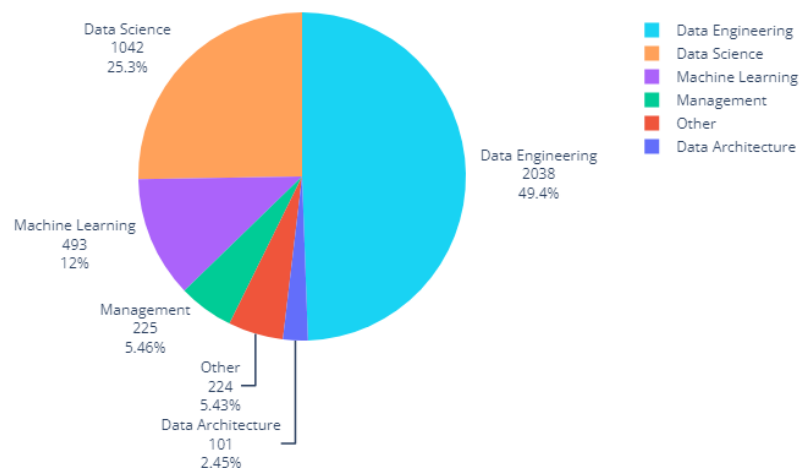
3.2.3. Phân tích mô tả

Phân tích mô tả trong phân tích dữ liệu là quá trình tóm tắt, mô tả và hiểu sâu về các đặc điểm, mẫu thái và thông tin quan trọng của tập dữ liệu. Với mục tiêu đó, ta sẽ tiến hành phân tích mô tả cho bộ dữ liệu của project theo cả 2 hướng phân tích đơn biến (trên từng biến) và phân tích đa biến (trên nhiều biến) bằng cách biểu diễn dưới các biểu đồ khác nhau.

- Biểu đồ 1: Biểu đồ phân bổ lĩnh vực làm việc

- + Dạng biểu đồ: Hình tròn (Pie chart)
- + Loại phân tích: Đơn biến ('job_role')
- + Kiểu dữ liệu: Phi số (object)

BIỂU ĐỒ HÌNH TRÒN PHÂN BỐ LĨNH VỰC LÀM VIỆC



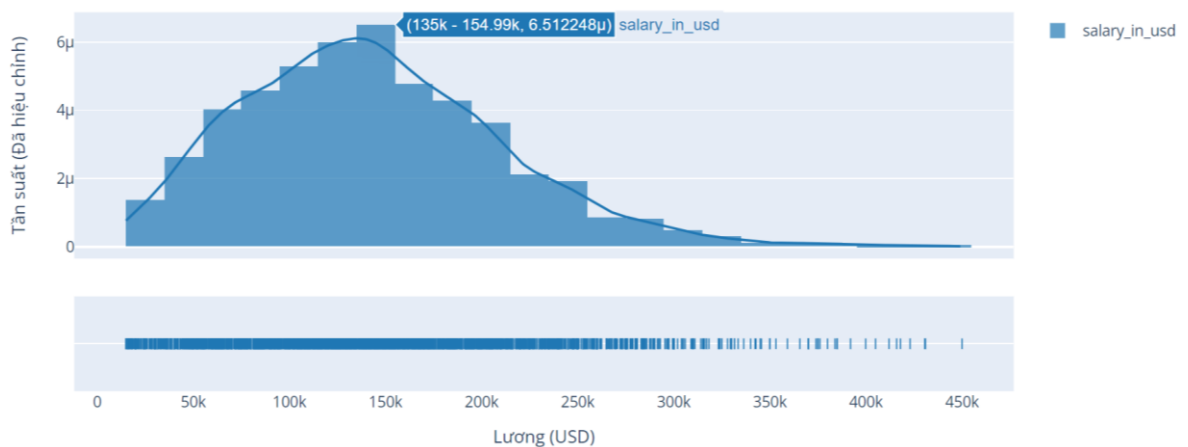
Hình 3.5. Biểu đồ hình tròn phân bổ lĩnh vực làm việc

Biểu đồ hình tròn là biểu đồ dùng để hiển thị tỷ lệ phần trăm của tần suất các giá trị duy nhất cụ thể. Như ở hình trên ta có thể nhìn thấy sự phân bố rõ ràng giữa các lĩnh vực công việc được phân tích từ cột dữ liệu ‘**job_role**’, ví dụ như mảng “Data Engineering” chiếm gần nửa các công việc khoa học dữ liệu hiện nay (49.4%).

- **Biểu đồ 2: Biểu đồ phân bố lương của công việc (tính theo USD)**

- + Dạng biểu đồ: Phân phối (Displot chart)
- + Loại phân tích: Đơn biến (‘salary_in_usd’)
- + Kiểu dữ liệu: Số nguyên (int64)

BIỂU ĐỒ DISPLOT CỦA LƯƠNG (USD)



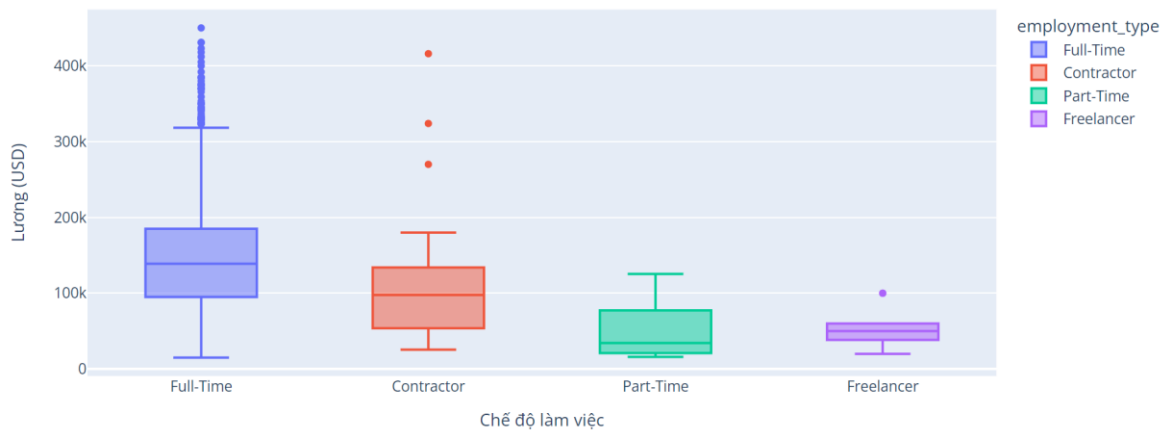
Hình 3.6. Biểu đồ phân phối lương tính theo USD

Displot chart thể hiện sự phân bố của một tập dữ liệu theo các khoảng giá trị. Biểu đồ là sự kết hợp giữa 3 thành phần (Histogram, KDE và Rug). Các khoảng giá trị được chia đều với khoảng cách lương là 20.000\$ cùng với tần suất được chuẩn hóa để tương thích với đường mật độ xác suất (KDE) của cột dữ liệu ‘**salary_in_usd**’. Nhìn vào biểu đồ, ta có thể thấy phân phối lương của ngành khoa học dữ liệu đang có xu hướng lệch dương (lệch về phía bên trái).

- **Biểu đồ 3: Biểu đồ phân bố lương (USD) theo chế độ làm việc**

- + Dạng biểu đồ: Hộp (Boxplot chart)
- + Loại phân tích: Đa biến (‘salary_in_usd’, ‘employment_type’)
- + Kiểu dữ liệu: Hỗn hợp (int64, object)

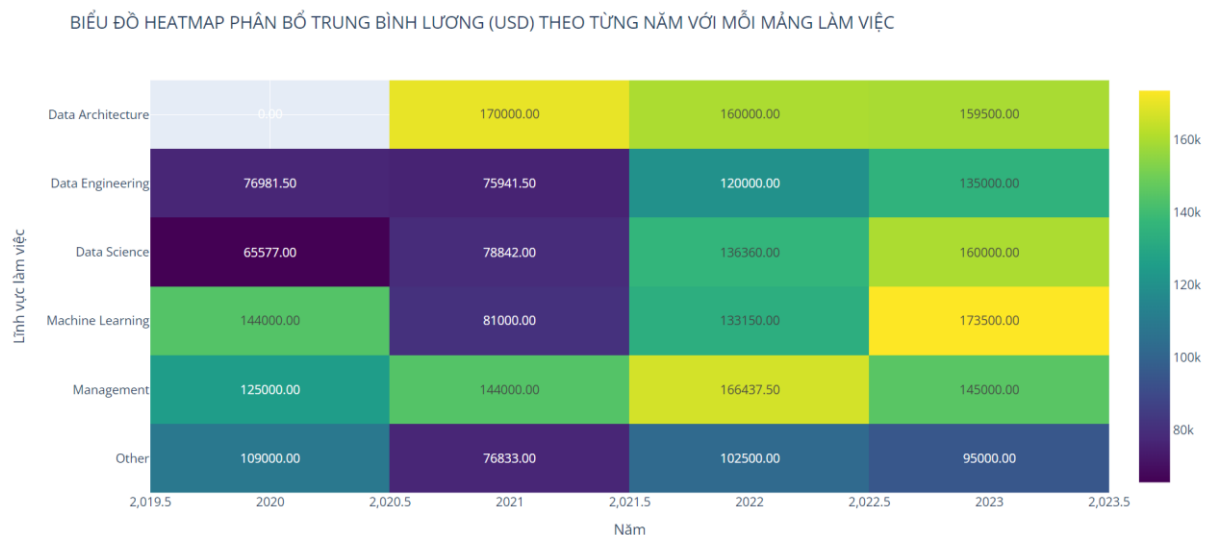
BIỂU ĐỒ BOXPLOT PHÂN BỐ LƯƠNG (USD) THEO CHẾ ĐỘ LÀM VIỆC



Hình 3.7. Biểu đồ Boxplot của lương tính theo USD và chế độ làm việc

Biểu đồ hộp thể hiện thống kê về phân phối của dữ liệu bằng cách hiển thị giá trị trung vị, khoảng tứ phân vị, và các giá trị ngoại lai. Bằng cách phân tích giá trị ‘salary_in_usd’ thông qua bộ lọc các chế độ làm việc khác nhau của ‘employment_type’, biểu đồ cho ta thấy rõ ràng chế độ làm việc toàn thời gian (Full-Time) đem lại lợi ích kinh tế tốt nhất so với tổng 4 chế độ làm việc.

- **Biểu đồ 4: Biểu đồ phân bố trung bình lương (USD) theo từng năm đối với mỗi lĩnh vực công việc**
 - + Dạng biểu đồ: Heatmap (Heatmap chart)
 - + Loại phân tích: Đa biến (‘salary_in_usd’ (mean), ‘employment_type’, ‘job_role’)
 - + Kiểu dữ liệu: Hỗn hợp (float64, object, object)

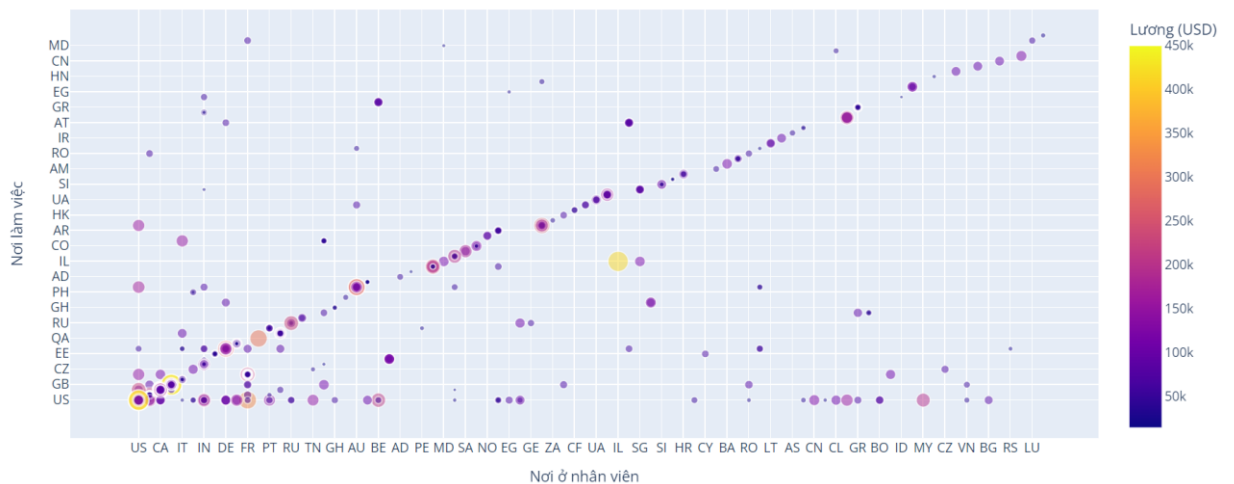


Hình 3.8. Biểu đồ Heatmap của trung bình lương (USD) qua từng năm với mỗi mảng làm việc

Biểu đồ heatmap sử dụng màu sắc để biểu thị mối quan hệ giữa hai chiều dữ liệu và hiển thị sự tương tác giữa các yếu tố trong một ma trận. Cụ thể với biểu đồ trên, hai chiều dữ liệu ở đây là ‘**work_year**’ và ‘**job_role**’, giá trị màu sắc là mức lương trung bình (mean của ‘**salary_in_usd**’). Qua quan sát ta có thể phân tích được lĩnh vực “Data Architecture” là một ngành mới khi mà ta chưa có dữ liệu về ngành này vào những năm 2020. Lĩnh vực này sau đó được ra mắt vào năm 2021 với mức lương trung bình ở mức cao (170.000\$) và có dấu hiệu hạ nhiệt nhẹ ở thời điểm hiện tại (năm 2023 với 159.500\$). Tương tự như thế có thể phân tích được xu hướng của các lĩnh vực còn lại theo thời gian.

- **Biểu đồ 5: Biểu đồ mức lương ảnh hưởng bởi nơi ở nhân viên và địa điểm công ty**
 - + Dạng biểu đồ: Phân tán (Scatter chart)
 - + Loại phân tích: Đa biến (‘salary_in_usd’, ‘employment_residence’, ‘company_location’)
 - + Kiểu dữ liệu: Hỗn hợp (float64, object, object)

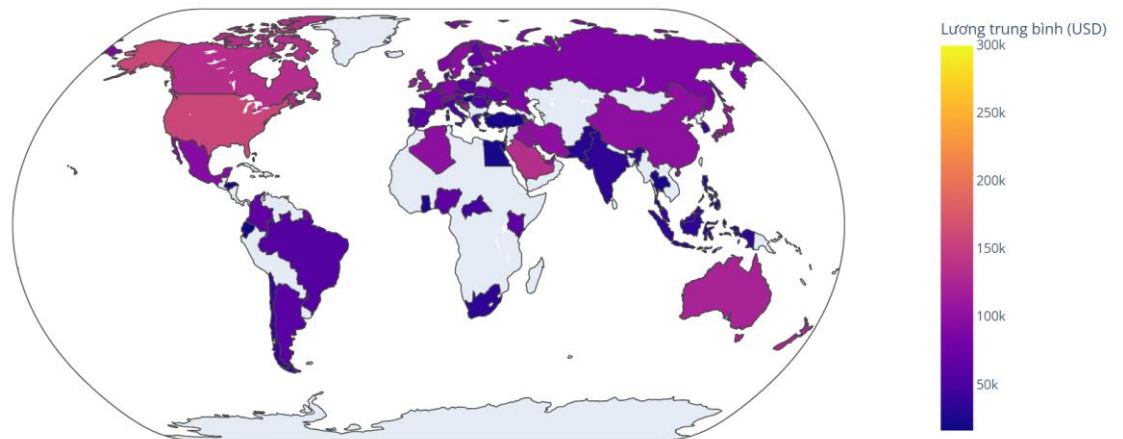
BIỂU ĐỒ MỨC LƯƠNG ẢNH HƯỞNG BỞI NƠI Ở NHÂN VIÊN VÀ ĐỊA ĐIỂM CÔNG TY



Hình 3.9. Biểu đồ phân tán mức lương ảnh hưởng bởi vị trí công ty và nơi ở nhân viên

Biểu đồ phân tán hiển thị các điểm dữ liệu trên một mặt phẳng, trong đó mỗi điểm biểu thị một cặp giá trị từ hai tập dữ liệu. Nó thường được sử dụng để tìm hiểu mối quan hệ hoặc xu hướng giữa các biến. Với biểu đồ trên, điểm dữ liệu được hiển thị trên trục hoành là **‘employment_resident’** (nơi ở nhân viên) và trục tung là **‘company_location’** (vị trí công ty). Kích thước của điểm dữ liệu có màu sắc và kích thước khác nhau dựa trên mức lương mà họ thu thập được thông qua giá trị **‘salary_in_usd’**. Bằng biểu đồ này, ta có thể phân tích và so sánh được sự chênh lệch về lương của nhân viên khi cùng một địa điểm làm việc nhưng khác nơi ở hoặc ngược lại.

- **Biểu đồ 6: Biểu đồ thể hiện mức lương trung bình ảnh hưởng bởi vị trí công ty**
 - + Dạng biểu đồ: Địa lý (Geographical chart)
 - + Loại phân tích: Đa biến (‘salary_in_usd’ (mean), ‘company_location’)
 - + Kiểu dữ liệu: Hỗn hợp (float64, object)



Hình 3.10. Biểu đồ địa lý thể hiện mức lương trung bình theo vị trí công ty

Biểu đồ địa lý là một loại biểu đồ đặc biệt khi sử dụng bản đồ để hiển thị dữ liệu dựa trên vị trí địa lý. Loại biểu đồ này thường được sử dụng để phân tích sự phân bố địa lý của dữ liệu hoặc hiển thị thông tin liên quan đến các vị trí cụ thể trên bản đồ. Với biểu đồ trên, bản đồ được tô màu dựa trên mức lương trung bình (mean) của **'salary_in_usd'** với mỗi vị trí tô dựa vào giá trị code quốc gia của **'company_location'** theo chuẩn ISO 3166. Từ biểu đồ ta có thể phân tích được vị trí địa lý tác động như nào đến mức lương của ngành Khoa học dữ liệu. Ví dụ như ở châu Mỹ, mức lương của Bắc Mỹ có xu hướng cao hơn so với ở Nam Mỹ.

3.2.4. Phân tích hồi quy

Với dự án hiện tại, mục tiêu được đặt ra là cần dự báo mức lương dựa theo mô hình hồi quy tuyến tính. Từ đó, ta đặt ra biến mục tiêu để dự báo (Target Value) chính là **'salary_in_usd'** của bộ dữ liệu.

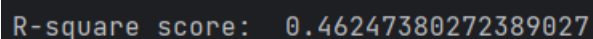
Trước tiên, chúng ta cần chất lượng dữ liệu để huấn luyện mô hình. Nhận thấy cột **'salary'** và cột **'salary_currency'** là không cần thiết do đã có cột **'salary_in_usd'** để chuyển đổi tất cả giá trị lương về cùng một tỷ giá (USD), vậy nên ta sẽ không đem 2 cột data này vào huấn luyện mô hình.

Để huấn luyện mô hình hồi quy tuyến tính, ta cần dữ liệu đầu vào hoàn toàn là dữ liệu số. Để làm được như vậy, ta cần phải sử dụng một thuật toán chuyển đổi dữ liệu phi số sang dữ liệu số phổ thông là **One-hot Encoding**.

Thuật toán này sẽ chuyển đổi tất cả các giá trị phi số trong cột dữ liệu thành các cột mới và hiển thị sự xuất hiện của nó dưới dạng giá trị nhị phân ('0' là không xuất hiện và '1' là có xuất hiện). Tuy nhiên thuật toán này có một nhược điểm là số lượng cột dữ liệu sinh ra sẽ trở nên nhiều phụ thuộc vào độ phức tạp của cột dữ liệu gốc, gây ảnh hưởng đến quá trình huấn luyện.

Sau khi đã xử lý cắt bỏ dữ liệu không cần thiết và chuyển đổi dữ liệu phù hợp, ta tiến hành huấn luyện cho mô hình hồi quy tuyến tính và đánh giá kết quả của thử nghiệm.

Để đánh giá kết quả thử nghiệm, chúng ta cần quan tâm đến R Square. Đây là một chỉ số được sử dụng trong việc đánh giá hiệu suất của mô hình hồi quy, thường được sử dụng để đo lường mức độ phù hợp của mô hình hồi quy với dữ liệu thực tế.



```
R-square score: 0.46247380272389027
```

Hình 3.11. Điểm R-square trên Terminal

Ở đây $R\text{ Square} = 0.46247380272389027$ nói rằng mô hình hồi quy tuyến tính này giải thích được ~46% sự biến thiên liên quan đến các biến độc lập, còn lại 54% là các yếu tố ngẫu nhiên khác. Để một mô hình hồi quy gọi là phù hợp thì R Square phải càng gần 100% càng tốt. Như vậy có thể đánh giá là mô hình hồi quy tuyến tính này có khả năng dự đoán trung bình thấp mức lương của ngành khoa học dữ liệu.

3.3. Đánh giá & Đề xuất

Phần phân tích mô tả đã phân tích được bộ dữ liệu ra các biểu đồ phù hợp và cho ta được cái nhìn tổng quan xoay quanh mức lương của ngành Khoa học dữ liệu trên toàn thế giới. Tuy nhiên, đối với bài toán dự báo, mô hình phân tích hồi quy tuyến tính đang không có điểm R Square tốt. Điều này cho ta thấy mô hình đang chưa thể nắm bắt tốt mối quan hệ giữa các đặc điểm và biến mục tiêu ('salary_in_usd').

Từ đó, ta có thể đề xuất một số phương án để cải thiện hiệu suất của mô hình như sau:

- Tạo các cột dữ liệu mới hoặc chuyển đổi các cột dữ liệu hiện có để nắm bắt tốt hơn mối quan hệ giữa các cột dữ liệu đầu vào và biến mục tiêu.
- Nghiên cứu rõ tầm quan trọng của từng cột dữ liệu và xem xét loại bỏ các dữ liệu không quan trọng để giảm nhiễu và tăng hiệu suất mô hình hồi quy.
- Thực hiện nhiều mô hình hồi quy khác để tìm ra được lựa chọn tối ưu cho mô hình của mình nhất (hồi quy Ridge, hồi quy Lasso, ...)
- Thử nghiệm với các thuật toán học máy khác, thay đổi các cách chuyển đổi dữ liệu hợp lý để xem có thể đem lại kết quả tốt hơn không.
- Thu thập thêm các cột dữ liệu mới liên kết chặt chẽ với biến mục tiêu nhằm mô hình có thêm nhiều thông tin tốt để huấn luyện và học hỏi.

3.4. Kết luận chương 3

Chương 3 đã trình bày phần thực nghiệm và đánh giá của dự án thông qua đầy đủ các bước từ tiền xử lý dữ liệu cho tới phân tích mô tả & bài toán dự báo. Từ đó đưa ra được các đánh giá và đề xuất phù hợp để cải thiện kết quả của dự án trong tương lai.

KẾT LUẬN

Trong bài tiểu luận này, chúng ta đã thực hiện phân tích dữ liệu và dự báo mức lương của ngành Khoa học dữ liệu bằng phương pháp hồi quy tuyến tính. Qua việc thu thập và xử lý dữ liệu liên quan đến các yếu tố như kinh nghiệm làm việc, tên công việc và vùng địa lý, chúng ta đã xây dựng một mô hình hồi quy tuyến tính để dự đoán mức lương trong ngành này. Kết quả đã cho thấy mô hình có khả năng dự báo mức lương với độ chính xác tương đối, tuy nhiên còn một số nhược điểm cần được xem xét.

Dựa trên kết quả phân tích và dự báo, chúng ta đưa ra một số kiến nghị nhằm cải thiện quá trình dự báo mức lương trong ngành Khoa học dữ liệu:

- **Tăng cường thu thập dữ liệu:** Cần tập trung vào việc thu thập dữ liệu chi tiết hơn về các yếu tố quan trọng ảnh hưởng đến mức lương.
- **Sử dụng mô hình phức tạp hơn:** Trong tương lai, nên xem xét sử dụng các phương pháp hồi quy phi tuyến hoặc mô hình học máy phức tạp hơn để đối phó tốt hơn với sự phức tạp và phi tuyến của dữ liệu.

Nghiên cứu về phân tích dữ liệu và dự báo mức lương trong ngành khoa học dữ liệu còn nhiều hướng phát triển tiềm năng:

- Mở rộng dữ liệu và phạm vi bằng cách thu thập dữ liệu từ nhiều ngành công nghiệp khác nhau, cùng với việc xem xét các yếu tố mới như tình hình kinh tế, xu hướng công nghệ, và sự biến đổi của thị trường lao động.
- Khám phá các phương pháp phân tích sâu hơn như phân tích chuỗi thời gian để tìm ra sự biến đổi của mức lương theo thời gian & các yếu tố ảnh hưởng.
- Kết hợp thêm nhiều dữ liệu địa lý để có thể xác định rõ ràng mối quan hệ giữa vị trí địa lý và thu nhập, từ đó đưa ra những nhận định hữu ích cho bộ phận quản lý nguồn nhân lực.

Tổng kết lại, nghiên cứu này đã tiến xa trong việc áp dụng phân tích dữ liệu và hồi quy tuyến tính để dự báo mức lương trong ngành Khoa học Dữ liệu. Tuy nhiên, vẫn còn nhiều cơ hội để mở rộng và phát triển đề tài trong tương lai.

TÀI LIỆU THAM KHẢO

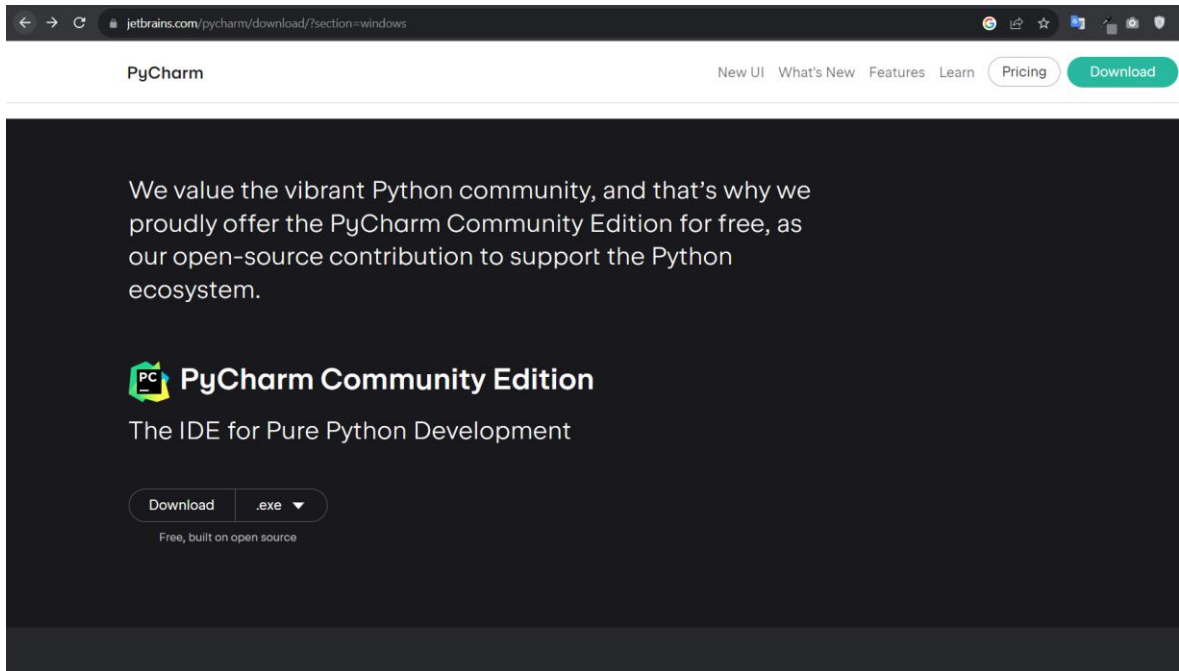
- [1] Solar Forecast Reconciliation and Effects of Improved Base Forecasts - <https://ieeexplore.ieee.org/document/8547846>
- [5] Get a full dataset of global AI, ML, Data salaries - <https://ai-jobs.net/salaries/download/>
- [3] Python Documentation - <https://docs.python.org/3/>
- [4] R Documentation - <https://cran.r-project.org/manuals.html>
- [2] “Data Analytics Made Accessible” Dr.Anil Maheshwari - NXB TechWorld

PHỤ LỤC

1. HƯỚNG DẪN SỬ DỤNG SẢN PHẨM

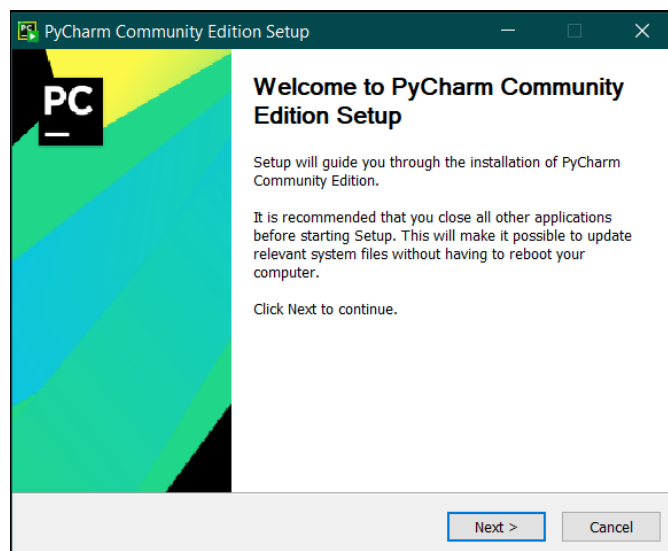
- Bước 1: Cài đặt và mở IDE PyCharm

Truy cập đường dẫn: <https://www.jetbrains.com/pycharm/download> và tiến hành tải xuống PyCharm (phiên bản Community).



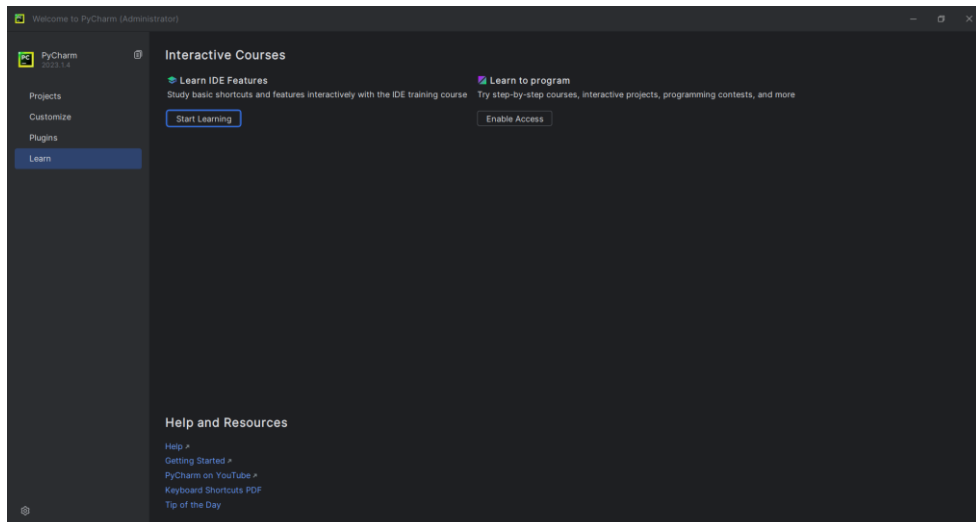
Hình 1. Tải xuống IDE PyCharm

Sau đó tiến hành cài đặt IDE PyCharm.



Hình 2. Cài đặt IDE PyCharm

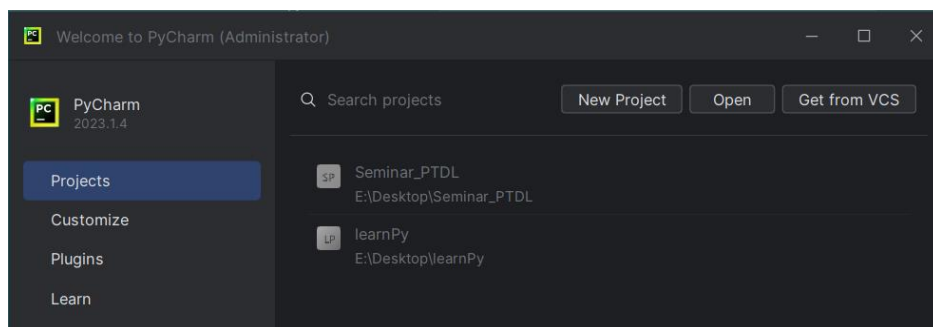
Sau khi cài đặt và mở IDE PyCharm, ta được giao diện như sau.



Hình 3. Giao diện IDE PyCharm

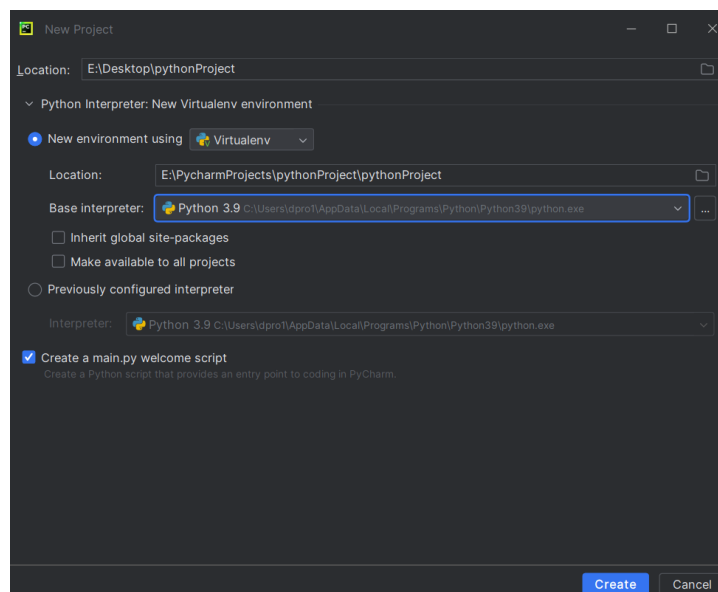
- Bước 2: Tạo project với IDE PyCharm.

Chọn Projects => New Project.



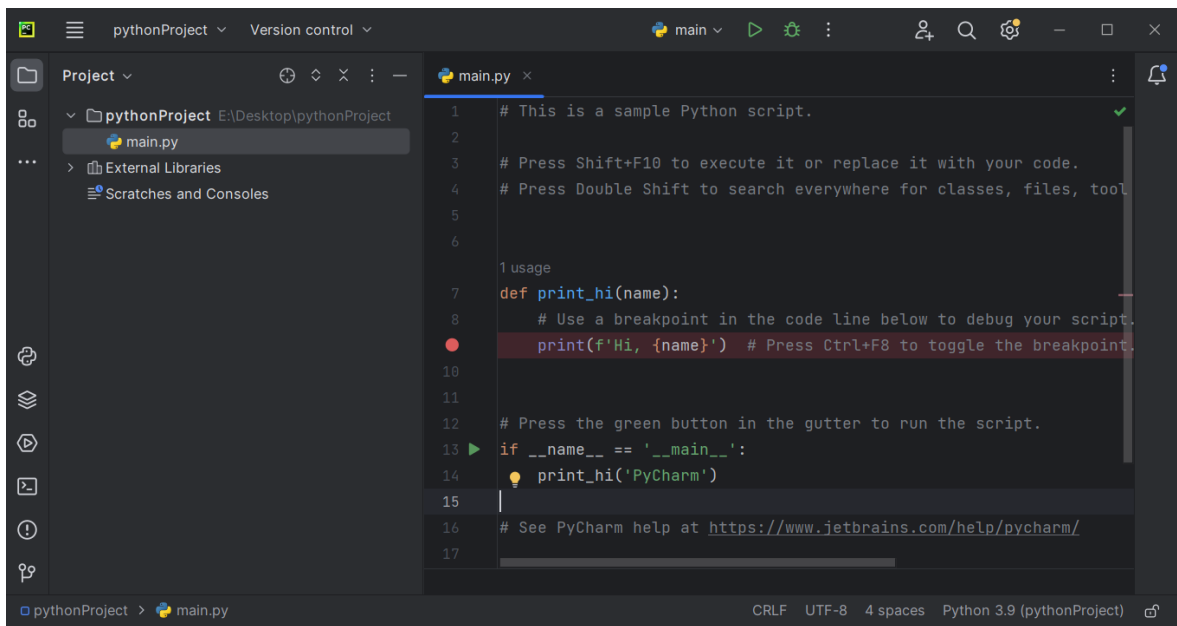
Hình 4. Khởi tạo Project

Lần lượt chọn môi trường lập trình (environment) và các đường dẫn lưu trữ, sau đó nhấn Create.



Hình 5. Thiết lập môi trường và chọn đường dẫn

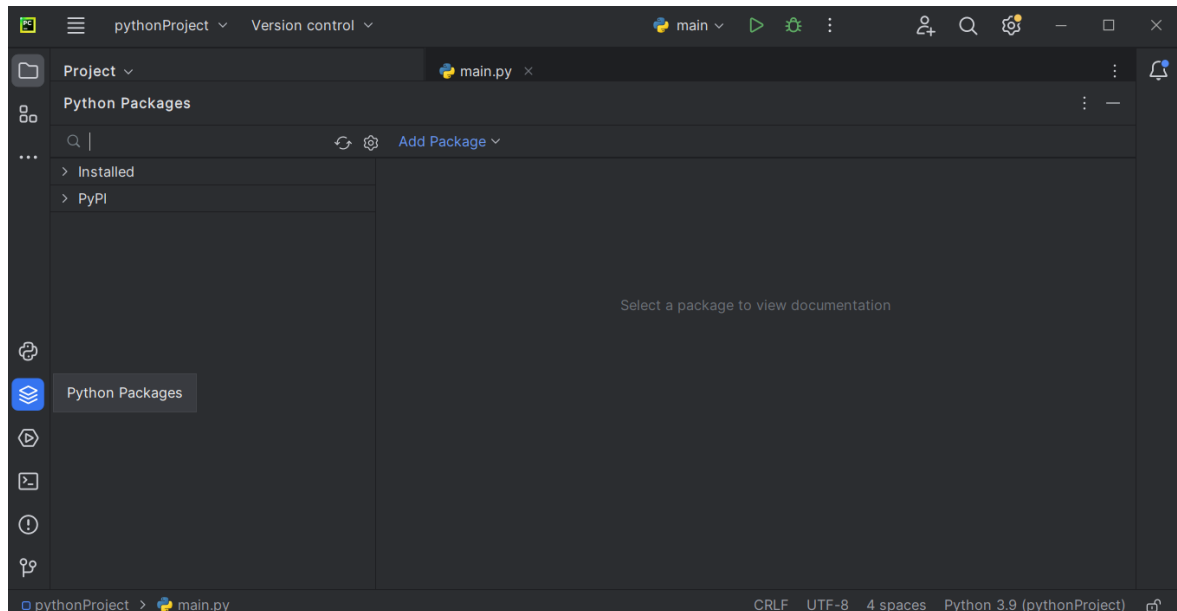
Sau khi thành công ta được giao diện như sau:



Hình 6. Giao diện lập trình của IDE PyCharm

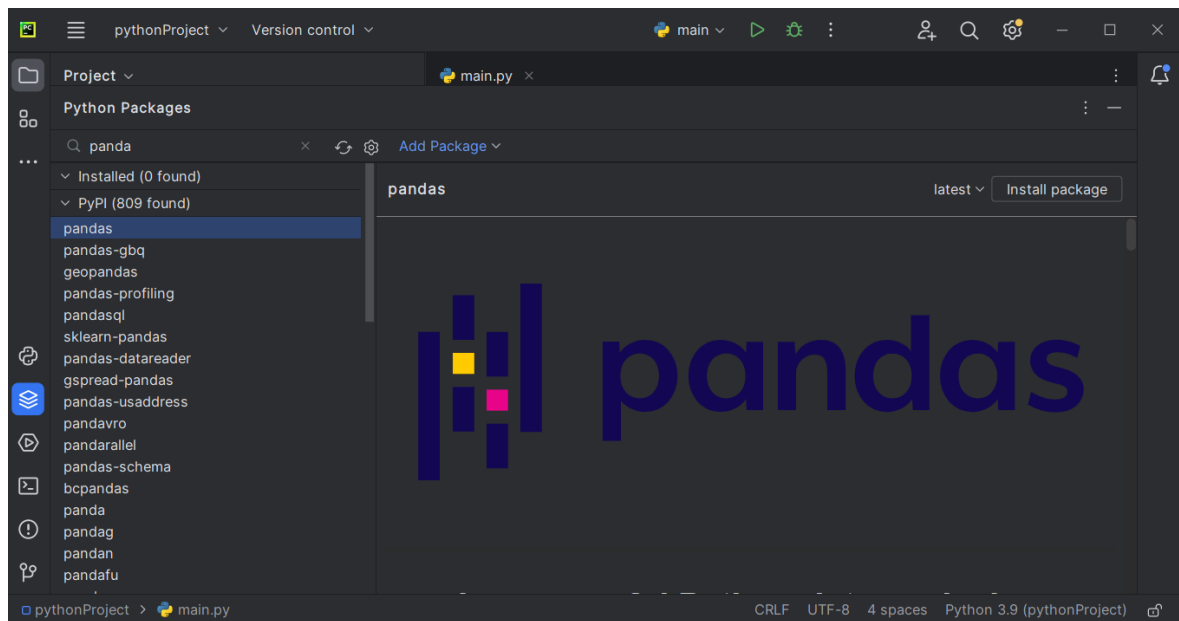
Bước 3: Tải xuống các Packages cần thiết

Ở bên trái giao diện, chọn Python Packages, tại đây ta có thể gõ tên các Packages cần thiết để cài đặt phục vụ giải quyết bài toán.



Hình 7. Giao diện lựa chọn tải xuống các Packages

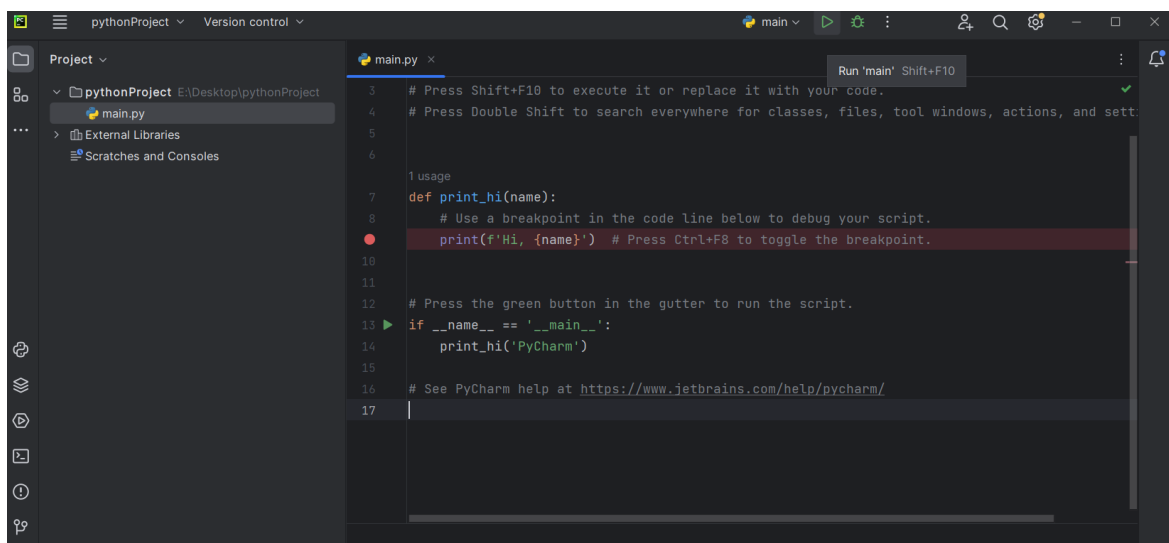
Gõ tên package cần cài đặt vào ô tìm kiếm, sau đó nhấn Install package để cài đặt. Các package cần cài đặt lần lượt là: pandas, numpy, scikit-learn.



Hình 8. Tải xuống và cài đặt các Package cần thiết

Bước 4: Tiến hành lập trình và chạy mã nguồn

Tiến hành lập trình bằng cách gõ mã nguồn vào file có tên mở rộng là .py, sau đó tiến hành chạy mã nguồn bằng nút Run.



Hình 9. Tiến hành lập trình và chạy mã nguồn

2. MÃ NGUỒN (SOURCE CODE)

```
import numpy as np
import pandas as pd
import plotly.express as px
import plotly.figure_factory as ff

import pycountry as pct

from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression

data = pd.read_csv('E:\Python\Data Science Salaries\salaries.csv')

#-----BẮT ĐẦU QUÁ TRÌNH TIỀN XỬ LÝ DỮ LIỆU CHO PHÂN TÍCH MÔ TẢ-----
-

# Tóm lược dữ liệu (Đo mức độ tập trung & mức độ phân tán)
description = data.describe()
mode = data.select_dtypes(include=['float64','int64']).mode().iloc[0]
mode.name = 'mode'
median = data.select_dtypes(include=['float64','int64']).median()
median.name = 'median'
description = description._append(mode)
description = description._append(median)
print(description)

# Kiểm tra tỷ lệ lỗi thiếu data
data_na = (data.isnull().sum() / len(data)) * 100
missing_data = pd.DataFrame({'Ty le thieu data': data_na})
print(missing_data)

# Kiểm tra data bị trùng
duplicated_rows_data = data.duplicated().sum()
print(f"\nSO LUONG DATA BI TRUNG LAP: {duplicated_rows_data}")
data = data.drop_duplicates()

# Quét qua các cột và đếm số lượng data riêng biệt
```

```

print("\nSO LUONG CAC DATA RIENG BIET:")
for column in data.columns:
    num_distinct_values = len(data[column].unique())
    print(f'{column}:{num_distinct_values} distinct values')

# Xem qua dataset
print(f'\n5 DONG DAU DATA SET:\n {data.head(5)}')

# Thay đổi giá trị để dataset dễ hiểu hơn
data['experience_level'] = data['experience_level'].replace({
    'SE': 'Senior level',
    'EN': 'Entry level',
    'EX': 'Executive level',
    'MI': 'Mid/Intermediate level'
})

data['employment_type'] = data['employment_type'].replace({
    'FL': 'Freelancer',
    'CT': 'Contractor',
    'FT': 'Full-Time',
    'PT': 'Part-Time'
})

data['company_size'] = data['company_size'].replace({
    'S': 'Small',
    'M': 'Medium',
    'L': "Large"
})

data['remote_ratio'] = data['remote_ratio'].astype(str) # Chuyển data về dạng chuỗi (ban đầu là dạng số)
data['remote_ratio'] = data['remote_ratio'].replace({
    '0': 'On-site',
    '50': 'Half-Remote',
    '100': 'Full-Remote'
})

```

```

# Định nghĩa hàm để gán lĩnh vực cho từng công việc
def assign_broader_category(job_title):
    data_engineering = ["Data Engineer", "Data Analyst", "Analytics Engineer", "BI Data Analyst",
                        "Business Data Analyst", "BI Developer", "BI Analyst", "Business Intelligence
Engineer",
                        "BI Data Engineer", "AI Engineer", "AI Research Engineer", "Azure Data
Engineer",
                        "BI Data Engineer",
                        "Big Data Engineer", "Cloud Data Engineer", "Cloud Database Engineer",
                        "Computer Vision Engineer",
                        "Computer Vision Software Engineer", "Consultant Data Engineer", "Data Analytics
Engineer",
                        "Data DevOps Engineer", "Data Engineer 2", "Data Infrastructure Engineer",
                        "Data Operations Engineer",
                        "Data Quality Engineer", "Data Science Engineer", "Data Visualization Engineer",
                        "Deep Learning Engineer",
                        "ETL Engineer", "Marketing Data Engineer", "ML Engineer", "MLOps Engineer",
                        "NLP Engineer", "Principal Data Engineer", "Research Engineer", "Software Data
Engineer"]
    data_scientist = ["Data Scientist", "Applied Scientist", "Research Scientist", "Deep Learning
Researcher",
                     "AI Scientist", "Applied Data Scientist",
                     "Decision Scientist", "Principal Data Scientist", "Staff Data Scientist"]
    machine_learning = ["Machine Learning Engineer", "ML Engineer", "Machine Learning
Developer",
                       "Applied Machine Learning Engineer",
                       "Machine Learning Engineer", "Machine Learning Infrastructure Engineer",
                       "Machine Learning Research Engineer",
                       "Machine Learning Software Engineer", "Principal Machine Learning Engineer",
                       "Staff Machine Learning Engineer",
                       "Machine Learning Researcher", "Principal Machine Learning Engineer",
                       "Applied Machine Learning Scientist",
                       "Machine Learning Scientist", "Head of Machine Learning", "Lead Machine
Learning Engineer"]
    data_architecture = ["Data Architect", "Big Data Architect", "Cloud Data Architect", "Principal
Data Architect",

```

```

        "AI Architect", "AWS Data Architect"]
management = ["Data Science Manager", "Director of Data Science", "Head of Data Science",
"Head of Data",
        "Data Lead", "Data Science Lead",
        "Data Scientist Lead", "Data Manager", "Data Operations Manager", "Data Analytics
Lead",
        "Data Science Tech Lead",
        "Lead Data Analyst", "Lead Data Engineer", "Lead Data Scientist"
        "Manager Data Management", "Data Analytics Manager",
        "Analytics Engineering Manager"]

```

```

if job_title in data_engineering:
    return "Data Engineering"
elif job_title in data_scientist:
    return "Data Science"
elif job_title in machine_learning:
    return "Machine Learning"
elif job_title in data_architecture:
    return "Data Architecture"
elif job_title in management:
    return "Management"
else:
    return "Other"

```

```

# Áp dụng hàm và tạo cột 'jobs_role'
data['job_role'] = data['job_title'].apply(assign_broader_category)
print(data) # Check lại dataset sau khi chuyển đổi dữ liệu ở terminal

```

```

#-----KẾT THÚC QUÁ TRÌNH TIỀN XỬ LÝ DỮ LIỆU CHO PHÂN TÍCH MÔ TẢ-----
--

```

```

#-----BẮT ĐẦU QUÁ TRÌNH PHÂN TÍCH MÔ TẢ-----

```

```

# Biểu đồ 1: Biểu đồ hình tròn phân bổ lĩnh vực làm việc (PTMT: Đơn biến - dữ liệu phi số)
job_role = data['job_role'].value_counts().sort_values(ascending=True)
fig1 = px.pie(values=job_role.values,
        names=job_role.index,

```

```

        color=job_role.index,
        title="BIỂU ĐỒ HÌNH TRÒN PHÂN BỐ LĨNH VỰC LÀM VIỆC")
fig1.update_traces(textinfo='label+percent+value',
                    textposition='outside')
fig1.show()

```

Biểu đồ 2: Biểu đồ displot dữ liệu lương tính theo USD (PTMT: Đơn biến - dữ liệu số)

```

fig2 = ff.create_distplot(hist_data=[data['salary_in_usd']],
                           group_labels=['salary_in_usd'],
                           bin_size=20000,
                           curve_type='kde')
fig2.update_layout(xaxis_title='Lương (USD)',
                    yaxis_title='Tần suất (Đã hiệu chỉnh)',
                    title='BIỂU ĐỒ DISPLOT CỦA LƯƠNG (USD)')
fig2.show()

```

Biểu đồ 3: Biểu đồ boxplot phân bố lương (USD) theo chế độ làm việc (PTMT: Đa biến (2) - dữ liệu hỗn hợp)

```

fig3 = px.box(data_frame=data,
               x = 'employment_type',
               y= 'salary_in_usd',
               color='employment_type',
               title='BIỂU ĐỒ BOXPLOT PHÂN BỐ LƯƠNG (USD) THEO CHẾ ĐỘ LÀM VIỆC')
fig3.update_layout(xaxis_title='Chế độ làm việc',
                    yaxis_title='Lương (USD)')
fig3.show()

```

Biểu đồ 4: Biểu đồ heatmap phân bố trung bình lương (USD) theo từng năm đối với mỗi mảng làm việc (PTMT: Đa biến (3) - dữ liệu hỗn hợp)

```

pivot_table = data.pivot_table(values='salary_in_usd',
                                index='job_role',
                                columns='work_year',
                                aggfunc='median')
fig4 = px.imshow(pivot_table,
                  labels=dict(x='Năm', y='Lĩnh vực làm việc'),
                  x=pivot_table.columns,
                  y=pivot_table.index,

```

```

        text_auto='.2f',
        color_continuous_scale='Viridis',
        title='BIỂU ĐỒ HEATMAP PHÂN BỐ TRUNG BÌNH LƯƠNG (USD) THEO TỪNG
        NĂM VỚI MỖI MẢNG LÀM VIỆC')
fig4.show()

```

Biểu đồ 5: Biểu đồ Scatter thể hiện mức lương ảnh hưởng bởi nơi ở của nhân viên và công ty (PTMT: Đa biến (3) - dữ liệu hỗn hợp)

```

fig5 = px.scatter(data_frame=data,
                  x='employee_residence',
                  y='company_location',
                  color='salary_in_usd',
                  size='salary_in_usd', opacity=0.5,
                  labels={'employee_residence': 'Nơi ở nhân viên', 'company_location': 'Nơi làm việc',
                          'salary_in_usd': 'Lương (USD)'},
                  hover_data=['salary_in_usd'],
                  title='BIỂU ĐỒ MỨC LƯƠNG ẢNH HƯỞNG BỞI NƠI Ở NHÂN VIÊN VÀ ĐỊA
        ĐIỂM CÔNG TY')
fig5.show()

```

Biểu đồ 6: Biểu đồ địa lý thể hiện mức lương trung bình theo vị trí công ty (PTMT: Đa biến (2) - dữ liệu hỗn hợp)

Hàm chuyển đổi code quốc gia thành tên quốc gia

def country_code_to_name(country_code): #Hàm 1

try:

return pct.countries.get(alpha_2=country_code).name # Nếu khả dụng trong thư viện thì chuyển về tên quốc gia

except:

return country_code #Còn không thì giữ nguyên để xử lý theo hàm 2

def country_code_to_name(code): #Hàm 2

try:

country = pct.countries.get(alpha_2=code)

return country.name

except:

return None

data['company_location'] = data['company_location'].apply(country_code_to_name)

```

avg_salary_by_location = pd.DataFrame(data.groupby('company_location',
as_index=False)['salary_in_usd'].mean())
fig6 = px.choropleth(data_frame=avg_salary_by_location,
                    locations='company_location',
                    locationmode='country names',
                    color='salary_in_usd',
                    hover_name='company_location',
                    color_continuous_scale='Plasma',
                    title='BIỂU ĐỒ ĐỊA LÝ THỂ HIỆN MỨC LƯƠNG TRUNG BÌNH THEO VỊ TRÍ
CÔNG TY',
                    labels={'salary_in_usd': 'Lương trung bình (USD)', 'company_location': 'Vị trí'},
                    projection='natural earth')
fig6.show()

```

#-----KẾT THÚC QUÁ TRÌNH PHÂN TÍCH MÔ TẢ-----

#-----BẮT ĐẦU QUÁ TRÌNH TIỀN XỬ LÝ DỮ LIỆU CHO PHÂN TÍCH HỒI QUY TUYẾN TÍNH-----

```

data = data[['work_year', 'experience_level', 'employment_type', 'job_title', 'salary_in_usd',
            'employee_residence', 'remote_ratio', 'company_location', 'company_size']] #Loại bỏ
'salary' và 'salary_currency'
labels_to_encode = ['experience_level', 'employment_type', 'job_title', 'employee_residence',
'remote_ratio', 'company_location', 'company_size'] #Các cột phi số
for label in labels_to_encode:
    data = data.join(pd.get_dummies(data[label], prefix=label))
    data.drop(label, axis=1, inplace=True)
X = pd.DataFrame(data.drop('salary_in_usd', axis=1))
Y = pd.DataFrame(data['salary_in_usd'])

```

#-----KẾT THÚC QUÁ TRÌNH TIỀN XỬ LÝ DỮ LIỆU CHO PHÂN TÍCH HỒI QUY TUYẾN TÍNH-----

#-----BẮT ĐẦU PHÂN TÍCH HỒI QUY TUYẾN TÍNH-----

```

model = LinearRegression()
model.fit(X,Y)

```



```
r2 = model.score(X,Y)
print("R-square score: ", r2)
```

```
#-----KẾT THÚC PHÂN TÍCH HỒI QUY TUYẾN TÍNH-----
```