

PHÂN TÍCH DỮ LIỆU - IE224.P11



PHÂN TÍCH VÀ DỰ ĐOÁN MỨC LƯƠNG NHÓM NGÀNH KHOA HỌC DỮ LIỆU

Nhóm 19

1. Nguyễn Phú Tài
2. Mai Văn Tân
3. Nguyễn Công Trúc
4. Trần Lê Nguyên Trung

Framework sử dụng

1. Jupyter Notebook, Python
2. Pandas, Matplotlib, Seaborn, Sklearn, Scipy
3. Power BI



Nội dung đồ án

MỞ ĐẦU

1. Giới thiệu

Tổng quan về đồ án, mục tiêu và tầm quan trọng của phân tích mức lương trong ngành khoa học dữ liệu

2. Mô tả bộ dữ liệu

Trình bày nguồn gốc, độ tin cậy, các đặc điểm chính của bộ dữ liệu, bao gồm các biến quan trọng và quy mô dữ liệu.

3. Phương pháp phân tích

Mô tả quy trình phân tích, các kỹ thuật và công cụ sử dụng để khai thác dữ liệu.

4. Phân tích thăm dò, sơ bộ

Tóm tắt các phát hiện ban đầu từ dữ liệu, như phân bố lương, xu hướng và mối quan hệ giữa các biến, ...

KẾT LUẬN

5. Kết quả phân tích

Trình bày các kết quả chính, mô hình dự đoán tốt nhất, và những đề xuất từ dữ liệu.

1. GIỚI THIỆU

PHÂN TÍCH VÀ DỰ ĐOÁN MỨC LƯƠNG NHÓM NGÀNH KHOA HỌC DỮ LIỆU

• Mục tiêu đề tài

Khám phá các yếu tố ảnh hưởng đến mức lương của công việc thuộc ngành Khoa học Dữ liệu, và xây dựng mô hình dự đoán phù hợp.

• Bộ dữ liệu

Bộ dữ liệu sử dụng là [Data Science Jobs & Salaries 2024](#), cung cấp bởi nguồn uy tín là Kaggle. Dữ liệu gồm thông tin tuyển dụng và mức lương của các vị trí trong ngành Khoa học Dữ liệu.

• Công cụ, thư viện, thuật toán

Python, Jupyter Notebook, Power BI; Pandas, Matplotlib, Seaborn, Sklearn, Scipy; Regressor Models, Cross-validation, RandomizedSearchCV.

Ghi chú: Toàn bộ đề tài và quá trình xử lý đều do nhóm tự thực hiện, đảm bảo không dựa trên nguồn nào khác.

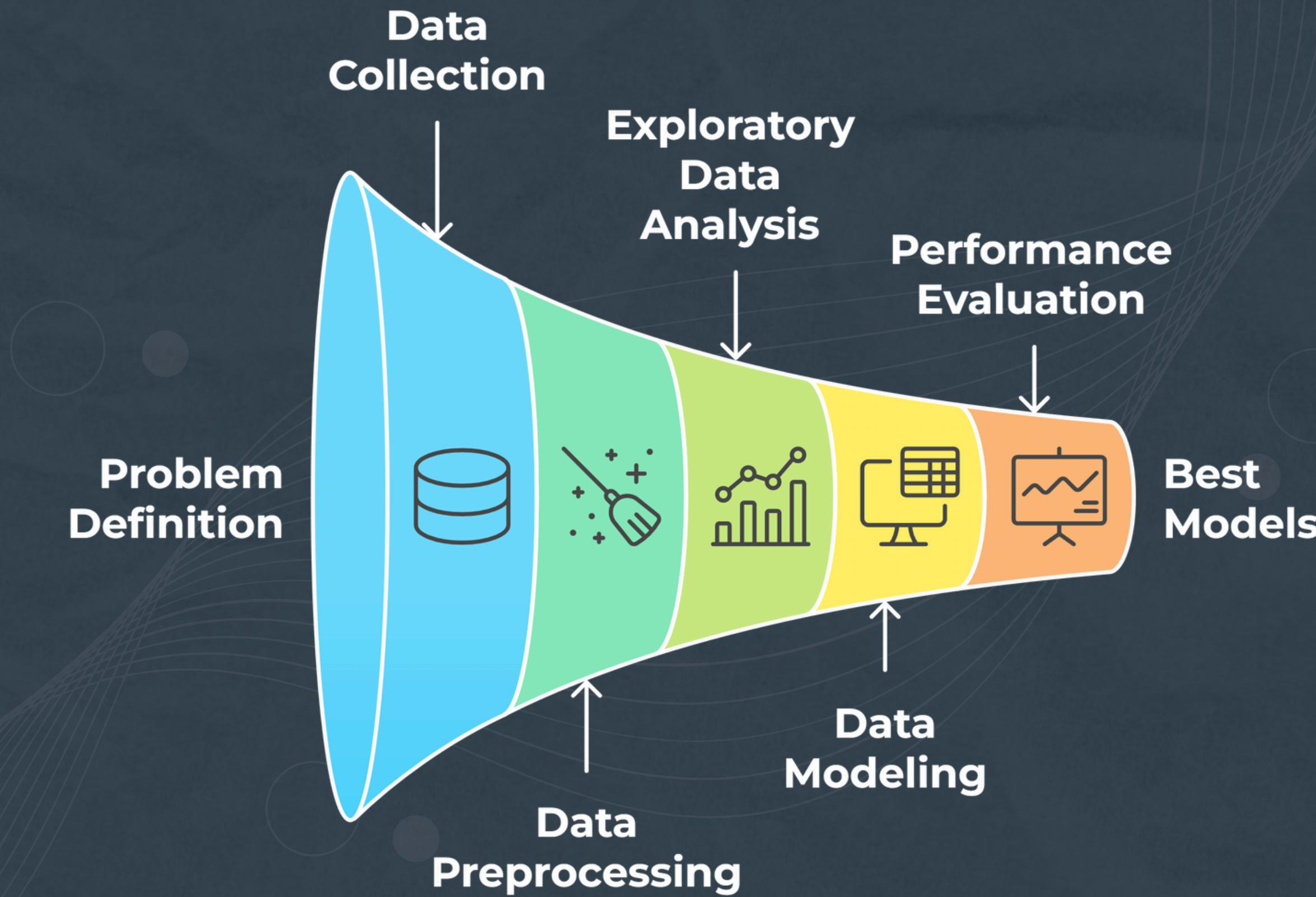
2. MÔ TẢ BỘ DỮ LIỆU

Tên thuộc tính	Mô tả	Kiểu dữ liệu	Khoảng giá trị
Job Title	Tên công việc	object (string)	“Data Scientist”, “Healthcare Data Scientist”, ...
Salary Estimate	Khoảng lương, đi kèm là ghi chú về cách tính lương hoặc thông tin này do ai cung cấp	object (string)	“\$53K-\$91K (Glassdoor est.)”, “\$21-\$34 Per Hour(Glassdoor est.)”, ...
Job Description	Mô tả công việc, có chứa các kỹ năng yêu cầu (chuỗi rất dài)	object (string)	“KnowBe4, Inc. is a high growth information sec...”, ...
Rating	Điểm số đánh giá về công ty	float64	[-1.0, 5.0]
Company Name	Tên công ty, kèm với điểm đánh giá	object (string)	“KnowBe4\n4.8”, “PNNL\n3.8”, ...
Location	Địa điểm làm việc	object (string)	“Linthicum, MD”, “Clearwater, FL”, ...

Ghi chú: Chúng tôi đảm bảo dữ liệu được chuyên gia từ Kaggle thu thập từ nguồn uy tín (glassdoor.com).

Tên thuộc tính	Mô tả	Kiểu dữ liệu	Khoảng giá trị
Headquarters	Địa điểm trụ sở chính	object (string)	“Baltimore, MD”, “Clearwater, FL”, ...
Size	Quy mô công ty	object (string)	“1 to 50 employees”, “51 to 200 employees”, ...
Founded	Năm thành lập	int64	[-1, 2019]
Type of ownership	Loại hình sở hữu công ty	object (string)	“Company - Private”, “Government”, ...
Industry	Ngành công nghiệp	object (string)	“Energy”, “Security Services”, ...
Sector	Lĩnh vực	object (string)	“Health Care”, “Business Services”, ...
Revenue	Doanh thu công ty	object (string)	“\$2 to \$5 billion (USD)”, “\$50 to \$100 million (USD)”, ...
Competitors	Tên các công ty đối thủ	object (string)	-1 hoặc “Novartis, Baxter, Pfizer”, “Travelers, Allstate, State Farm”, ...

3. PHƯƠNG PHÁP PHÂN TÍCH



Data Preprocessing

Data Preprocessing

Quan sát

Missing Values, Outliers

Tồn tại giá trị khuyết hoặc không rõ ý nghĩa, sẽ được xử lý bằng cách điền khuyết hoặc tái định nghĩa cho phù hợp với mục đích phân tích.

Tồn tại nhiều outlier.

Ví dụ: Các giá trị -1 trong cột “Rating”, “Industry”, ...

Duplicates Rows

Có 356/956 dòng bị lặp (duplicated), đây là những nội dung tuyển dụng được đăng từ lần thứ 2 trở đi. Có 2 hướng xử lý:

1. Giữ lại để phân tích xu hướng tuyển dụng của thị trường.
2. Xóa để xây dựng mô hình dự đoán lương nhằm tránh việc mô hình thiên vị (biased).

Data Types, Statistics

Các biến đều có kiểu dữ liệu phù hợp, không bị lỗi kiểu trong quá trình chuyển đổi.

Kiểm tra giá trị của từng biến và phân phối thống kê (mean, std, min, max) của chúng.

Data Preprocessing

Làm sạch

2. Chuẩn hóa tên công việc

Tạo cột “job_simplified” để đơn giản hóa và gom nhóm, cột “seniority” để trích xuất cấp bậc tuyển dụng từ tên công việc.

4. Trích xuất kỹ năng yêu cầu

Trích xuất kỹ năng được yêu cầu nhiều từ “Job Description” và tạo các đặc trưng tương ứng.

1. Loại bỏ cột thừa

Loại bỏ cột thừa hoặc không cần thiết cho mục đích: “Unnamed: 0”, “Competitors”.

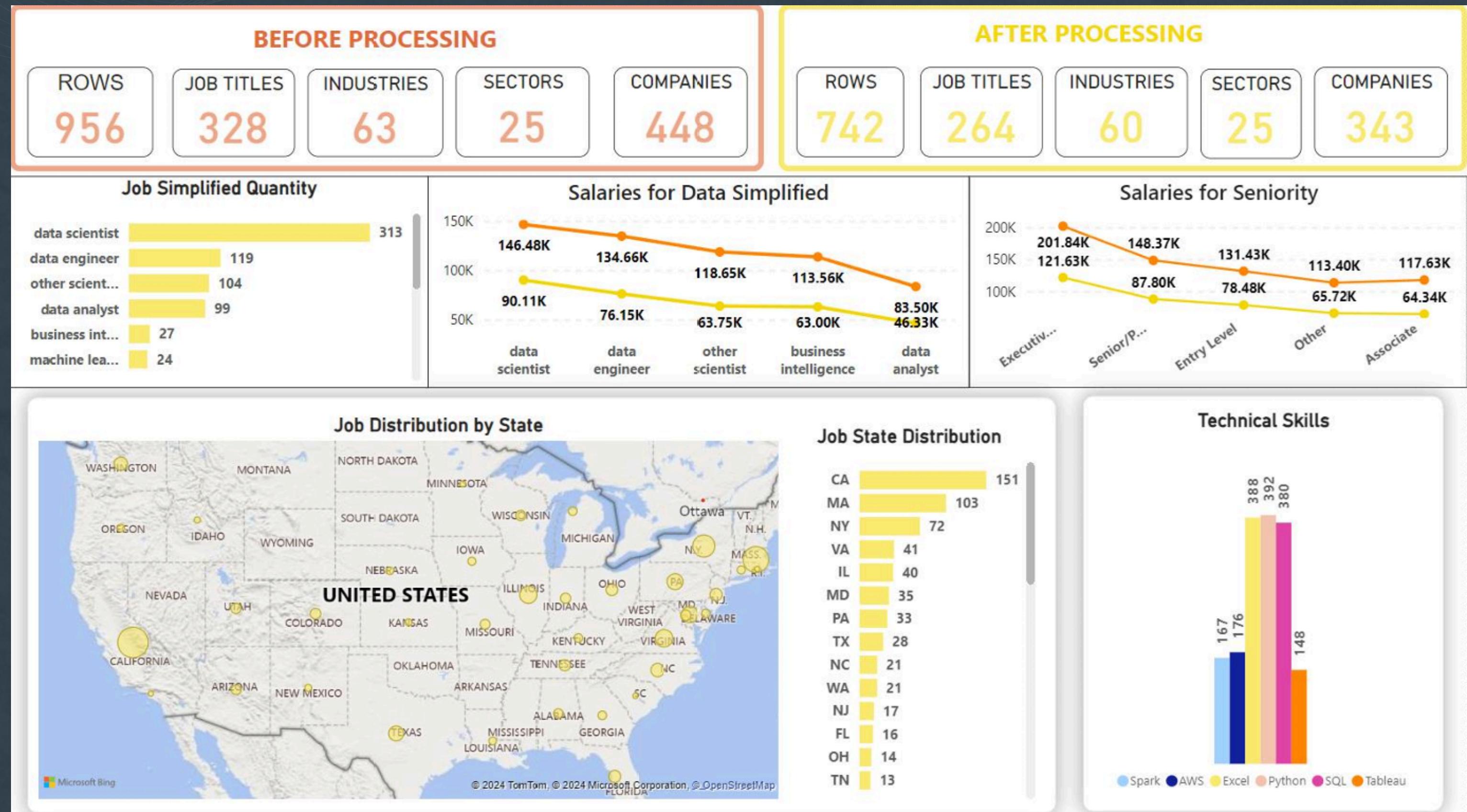
3. Chuẩn hóa mức lương

Loại bỏ các dòng có giá trị -1.
Tạo cột “Hourly” và “employer_provided” để đánh dấu công việc theo giờ và lương do nhà tuyển dụng cung cấp.
Làm sạch cột “Salary Estimate” và tạo các cột “Min Salary”, “Max Salary”, “Average Salary”.

5. Chuẩn hóa thông tin công ty

Tạo cột “Rating Category” để phân nhóm từ “Rating”.
Thay giá trị khuyết trong “Size”, “Revenue”, “Type of ownership” bằng “Unknown”.
Làm sạch “Company Name” và thêm cột “Age” từ “Founded” để thể hiện tuổi công ty.

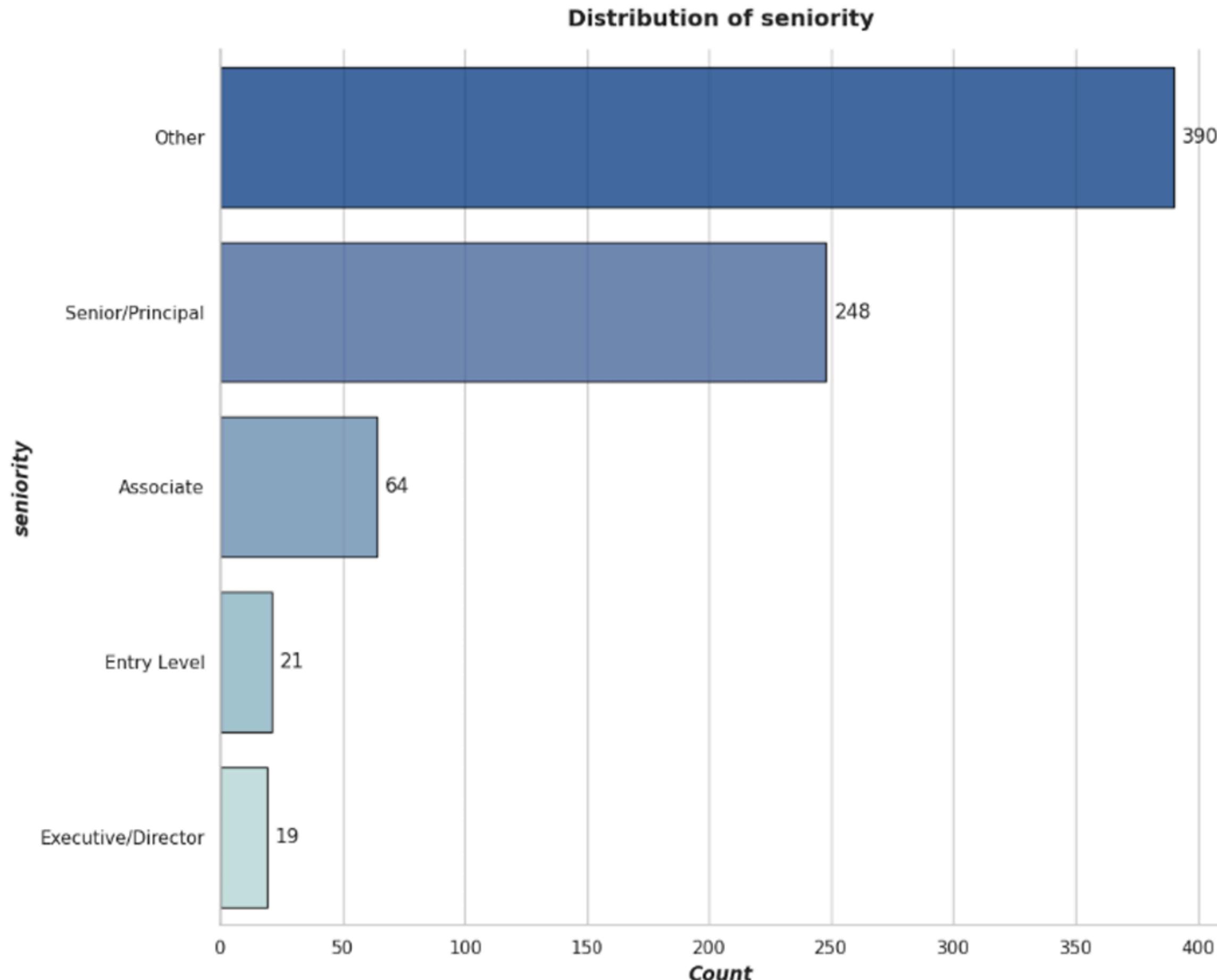
4. PHÂN TÍCH THĂM DÒ/SƠ BỘ



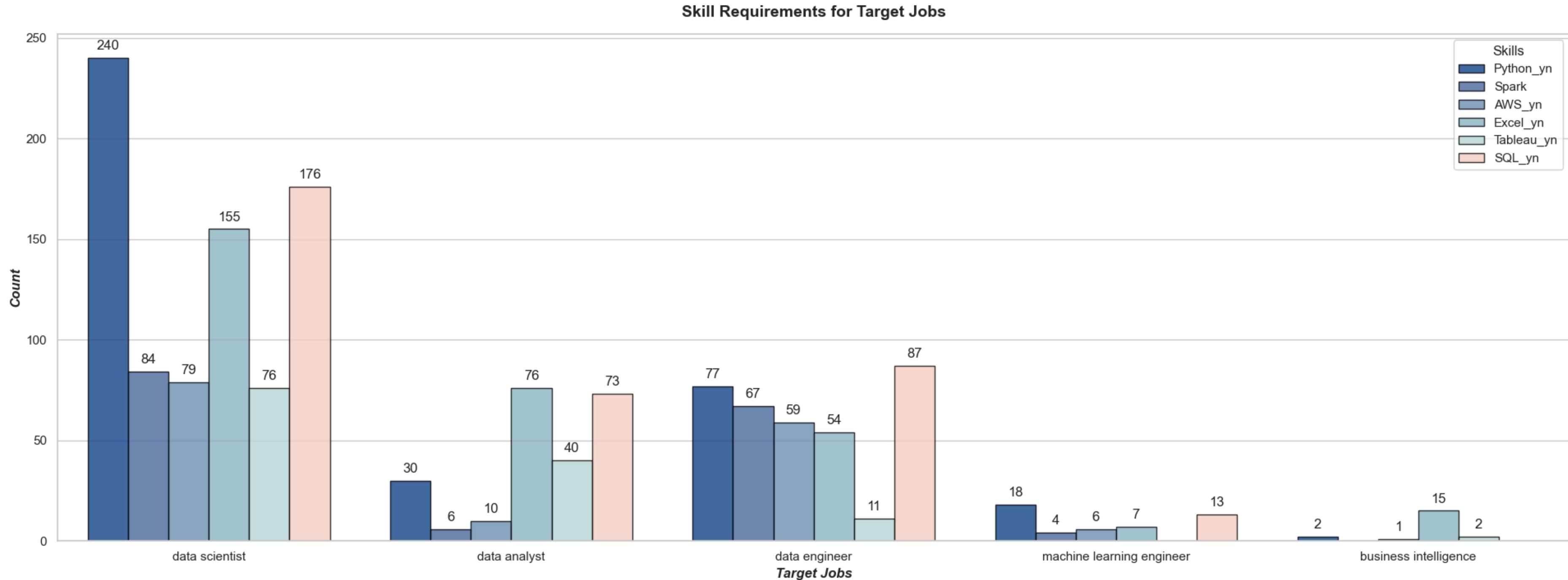
5. KẾT QUẢ PHÂN TÍCH

Exploratory Data Analysis

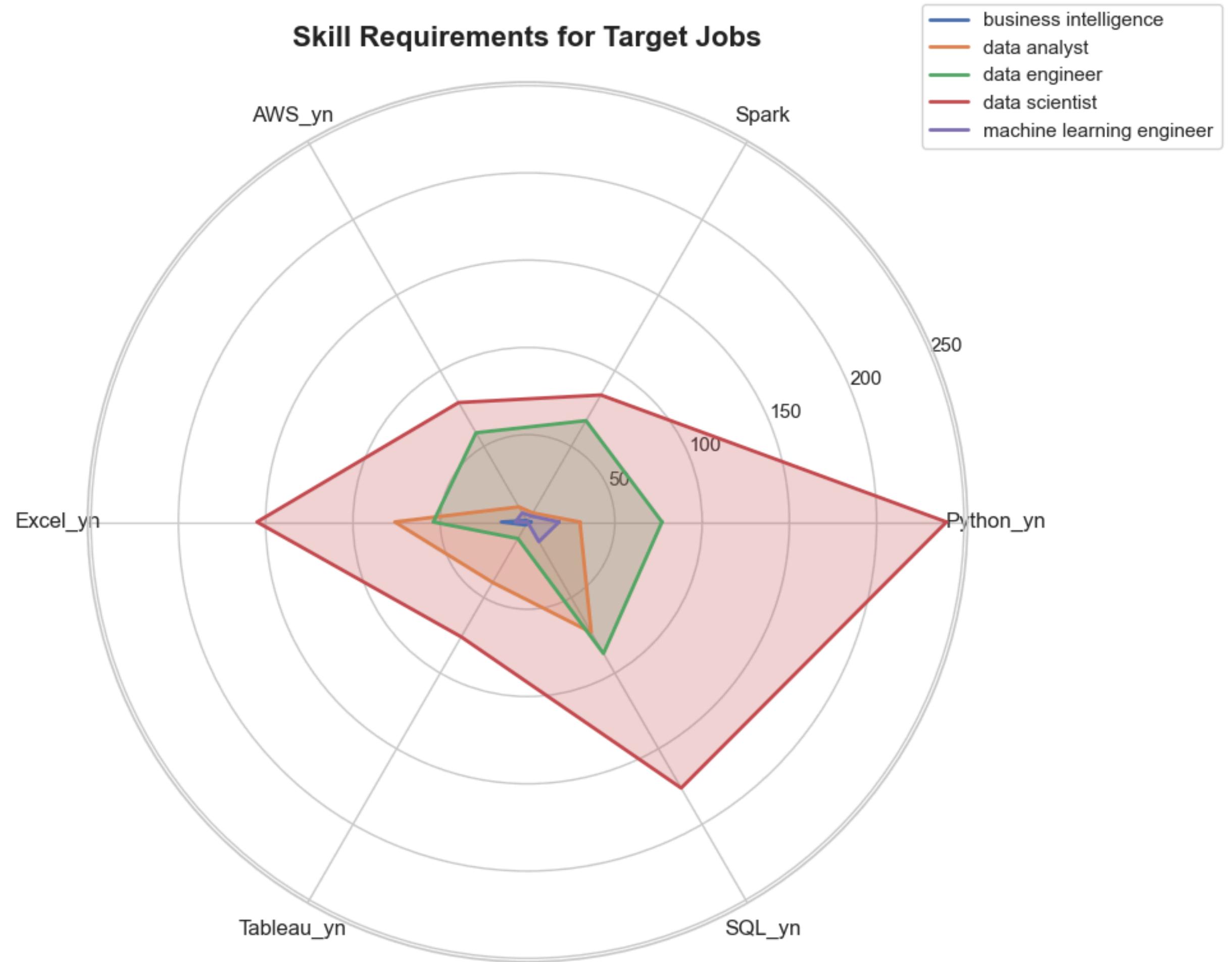
1. Thông tin về công việc



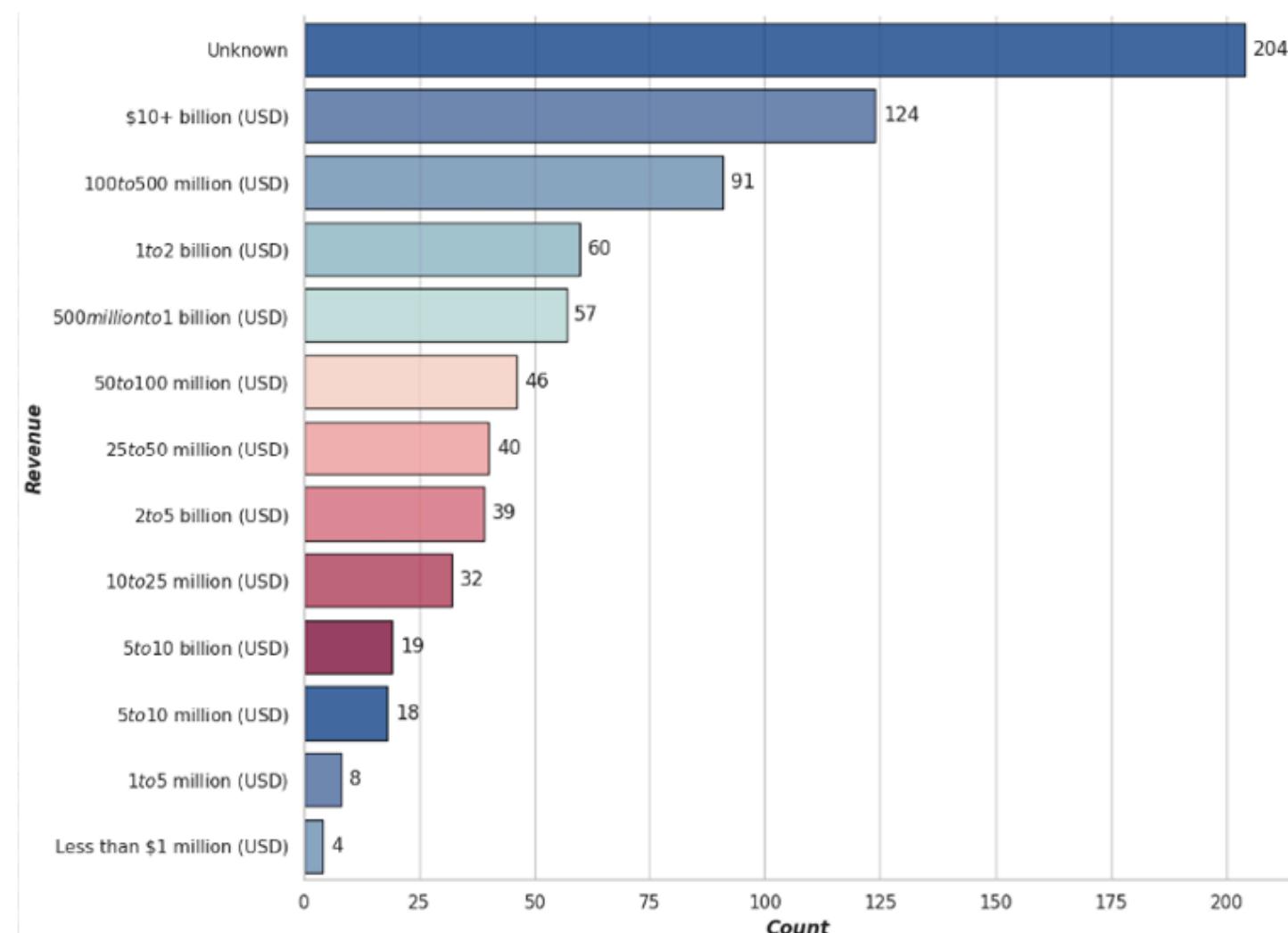
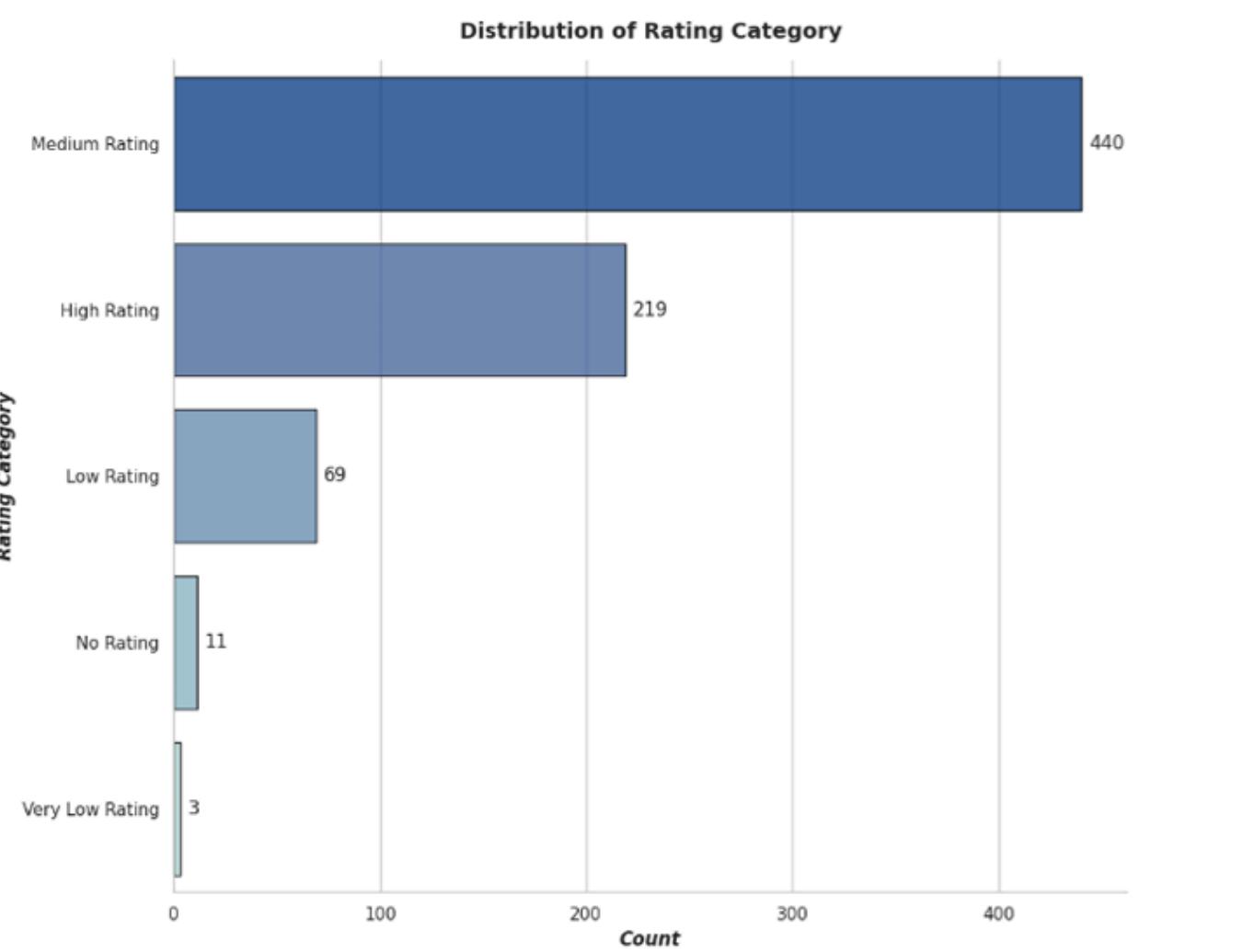
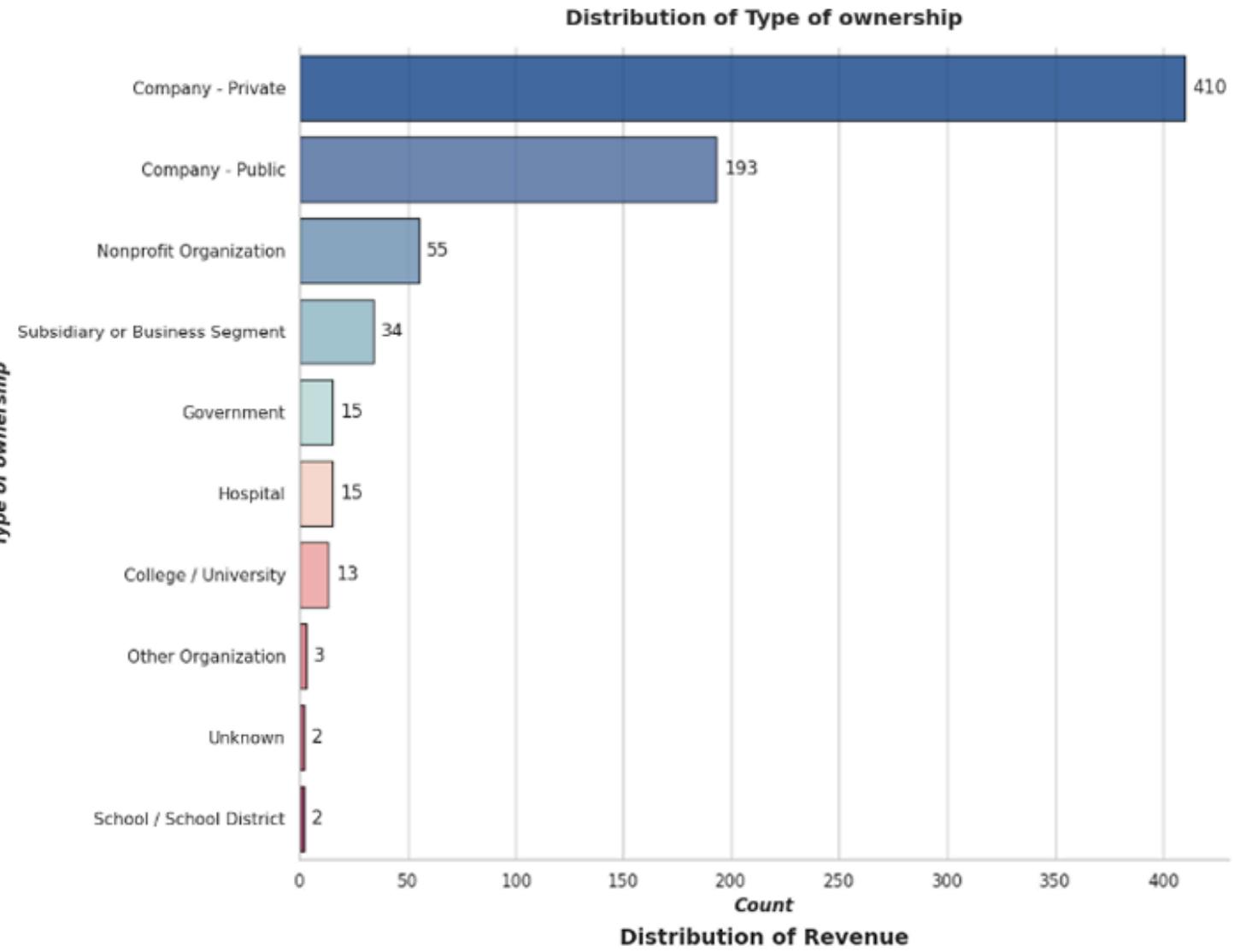
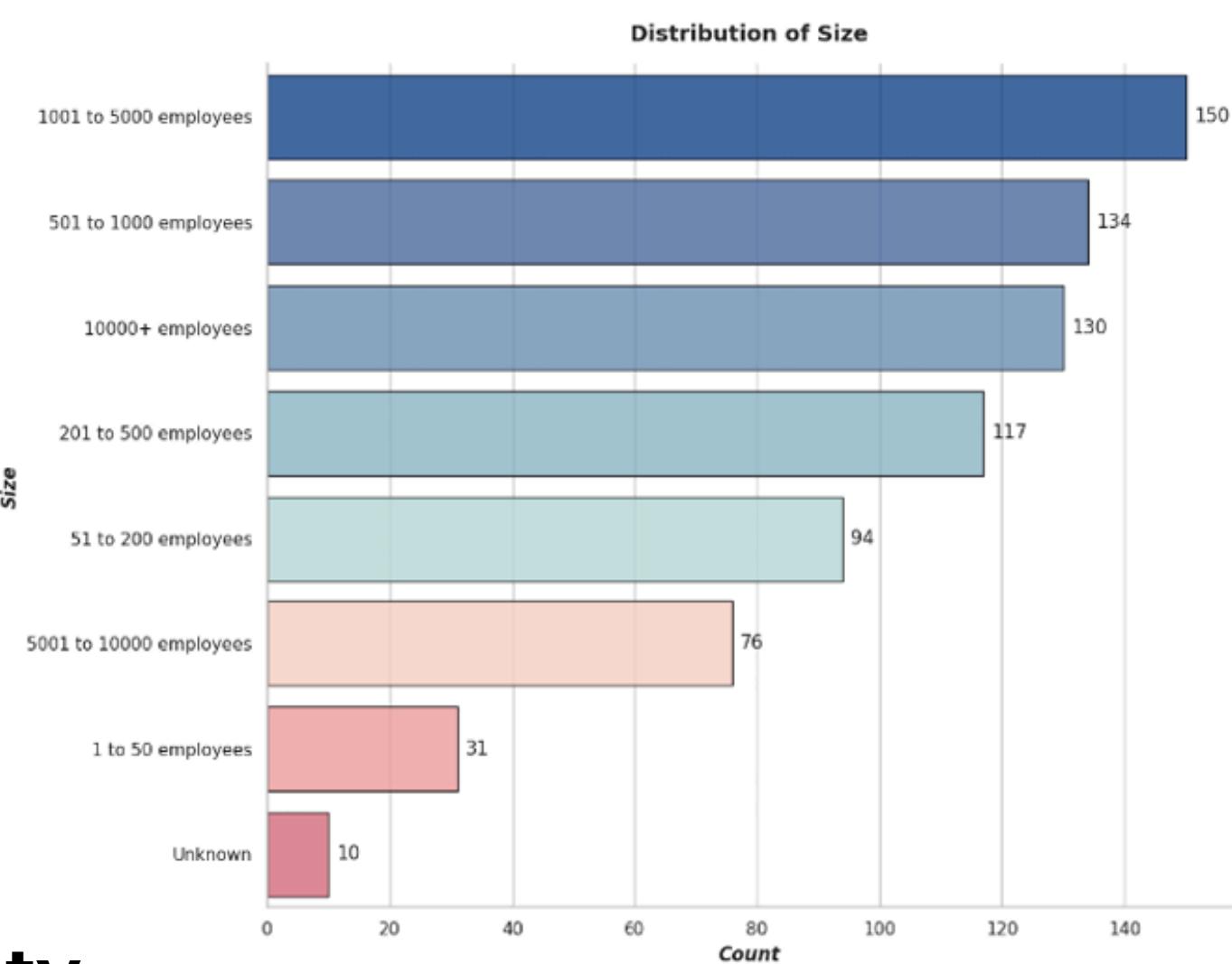
2. Kỹ năng cần thiết



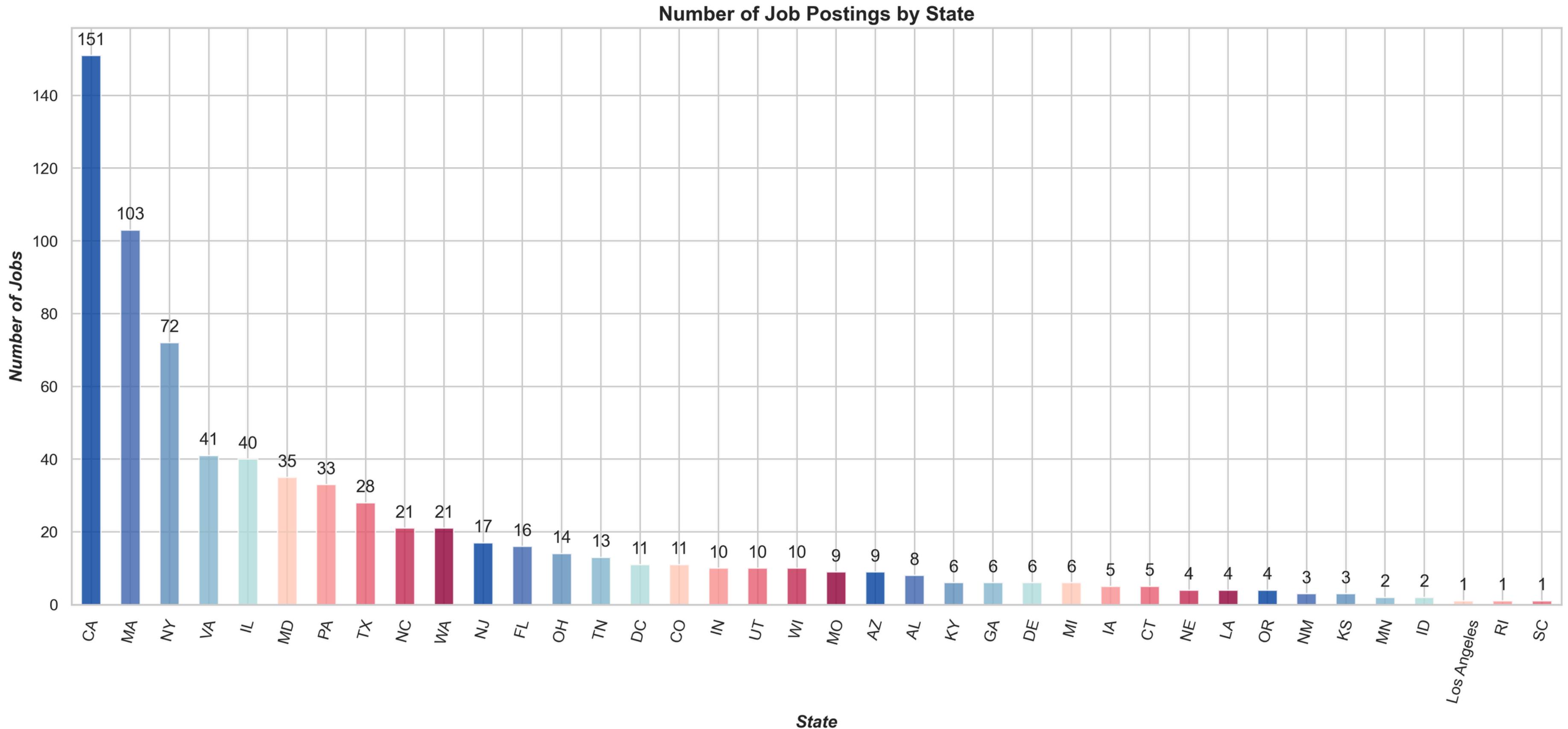
2. Kỹ năng cần thiết (Góc nhìn khác)



3. Thông tin về công ty



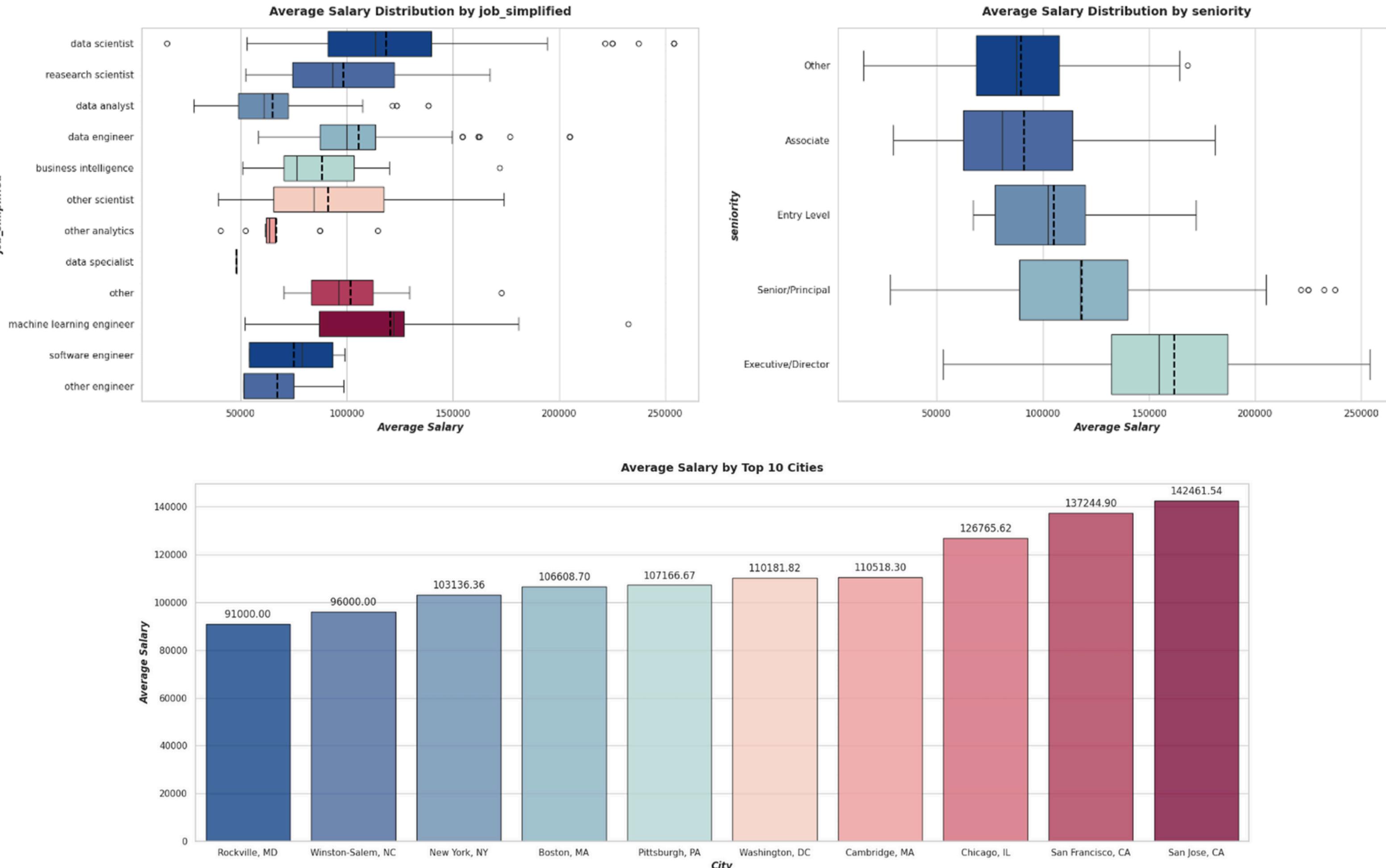
4. Thông tin về địa điểm và khu vực



Ảnh hưởng của kinh nghiệm làm việc

Ảnh hưởng của địa lý vị trí

5. Thông tin về mức lương trên thị trường



Một số nhận xét

Phân tích nhóm công việc

Giúp nhóm các công việc tương tự vào cùng một nhóm như Data Scientist, Data Engineer, Data Analyst,... Điều này giúp chúng ta hiểu về xu hướng nghề nghiệp và nhu cầu thị trường trong lĩnh vực dữ liệu.

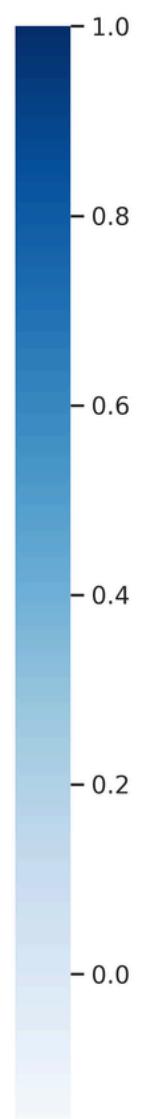
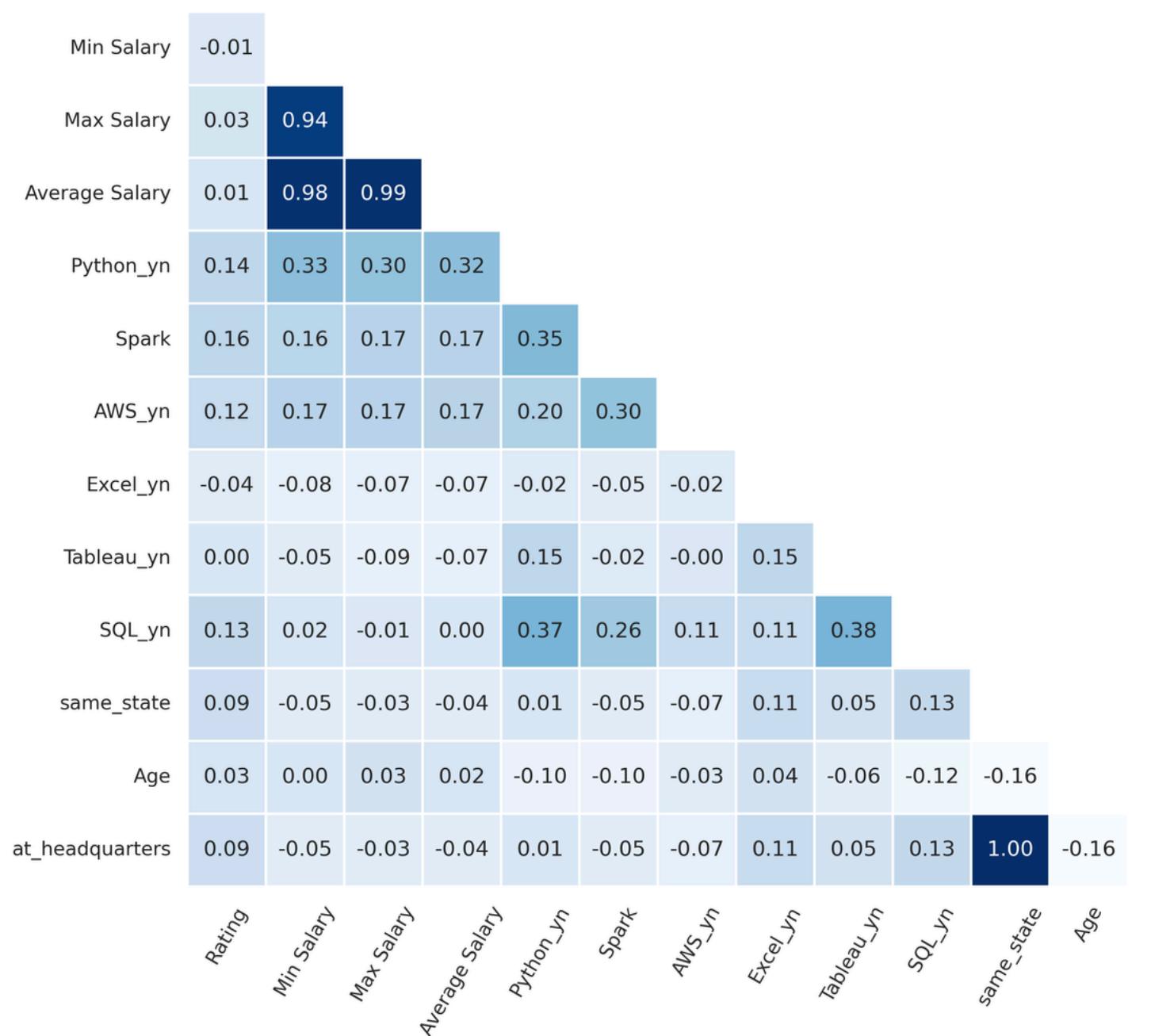
Phân tích dữ liệu địa lý

Cho thấy các bang như California và Massachusetts có cơ hội việc làm cao, phản ánh sự tập trung công việc tại các trung tâm công nghệ và khoa học như Silicon Valley, Boston.

Phân tích dữ liệu kỹ năng

Nhận thấy những kỹ năng như Python, SQL, và Excel là bắt buộc, cho thấy nhu cầu về phân tích dữ liệu và lập trình là cốt lõi.

Feature Selection



	F-statistic	p-value	Significance
Company Name	4.740723	4.549181e-48	True
job_simplified	24.894824	6.405628e-44	True
seniority	43.892767	4.572969e-33	True
Headquarters	3.234980	5.036396e-27	True
Location	2.978916	1.386686e-23	True
job_state	4.248494	9.027568e-15	True
Industry	3.265136	1.173209e-13	True
Type of ownership	7.725791	6.611024e-11	True
Sector	4.283655	7.908559e-11	True
Revenue	5.835820	1.011894e-09	True
Size	5.868690	1.200392e-06	True
Rating Category	5.663335	1.712989e-04	True

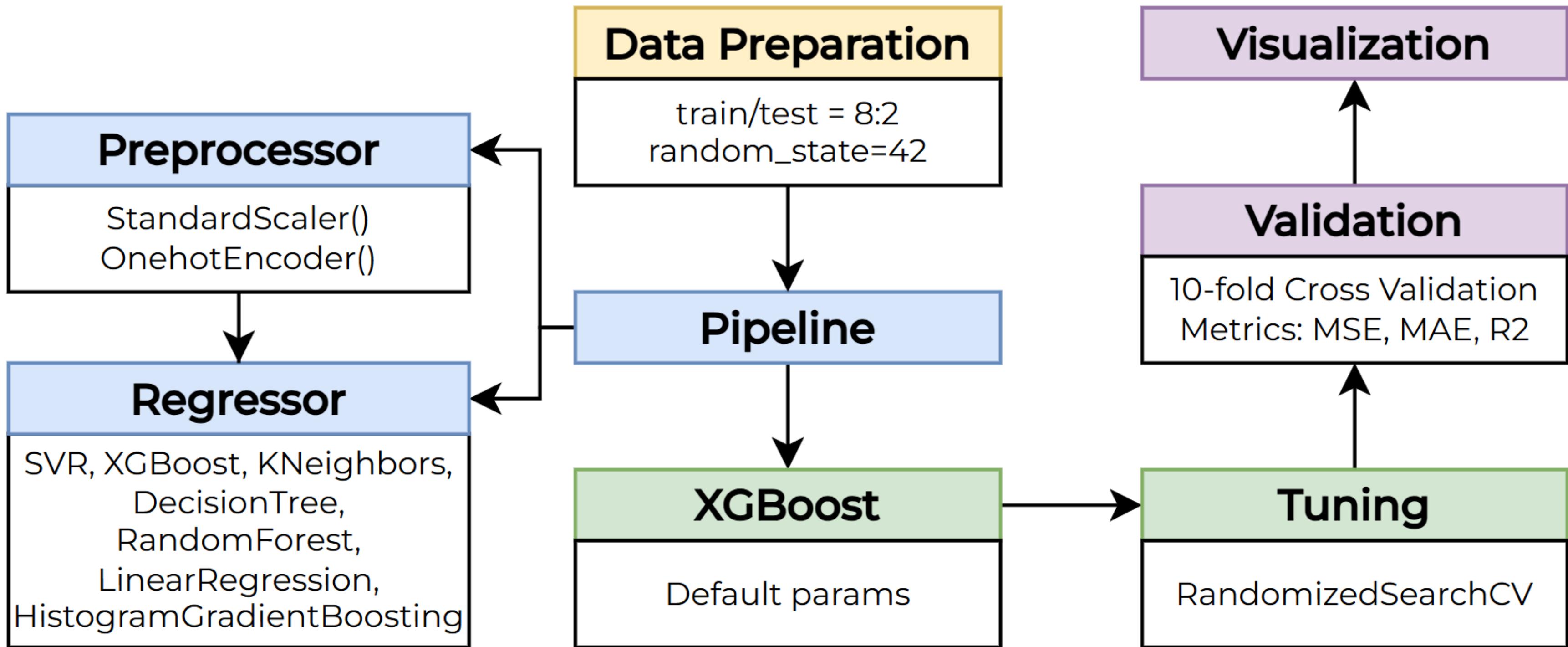
Đặc trưng quan trọng

Correlation Matrix: Python_yn, Spark, AWS_yn, Average Salary

Phân tích ANOVA: Company Name, job_simplified, seniority, Headquarters, Location, job_state, Industry, Type of ownership, Sector, Revenue, Size, Rating Category

Data Modeling & Performance Evaluation

Data Modeling - Training & Evaluation



Data Modeling - Training

Training

	Mô hình	MSE	MAE	R2
0	XGBoost	2.687e+08	1.065e+04	8.286e-01
1	RandomForest	2.936e+08	1.127e+04	8.127e-01
2	HistogramGradientBoosting	4.009e+08	1.456e+04	7.443e-01
3	DecisionTree	4.605e+08	1.024e+04	7.062e-01
4	KNeighbors	8.300e+08	2.173e+04	4.706e-01
5	SVR	1.616e+09	3.117e+04	-3.078e-02
6	LinearRegression	2.511e+33	1.946e+16	-1.602e+24

XGBoost với tham số mặc định

Mean Squared Error (MSE): 268723315.569

Mean Absolute Error (MAE): 10650.671

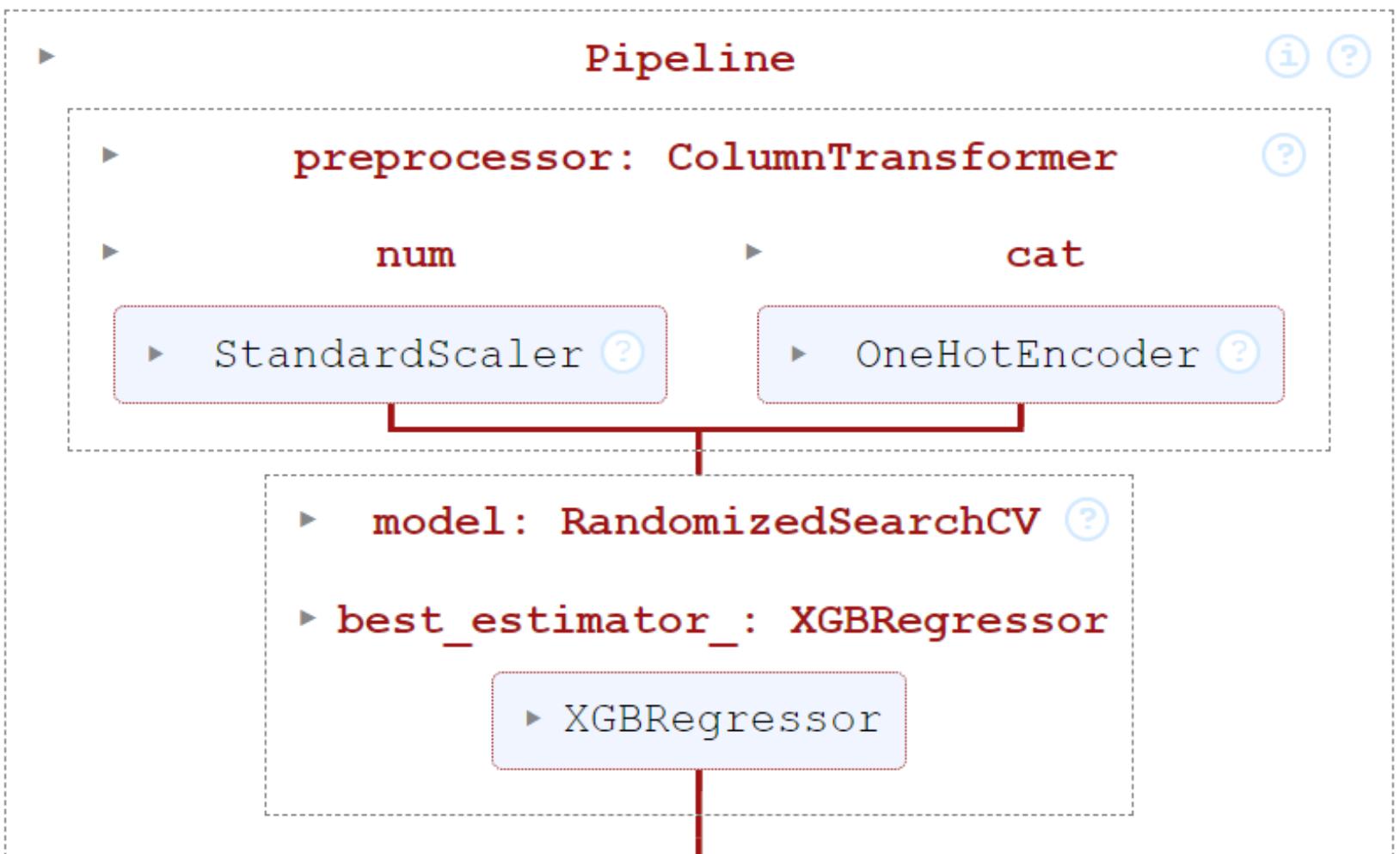
R-squared (R2) Score: 0.829

XGBoost có hiệu suất
tốt nhất

Data Modeling - Tuning

RandomizedSearchCV

Tham số tốt nhất



	Value
reg_lambda	1.000e-05
reg_alpha	1.000e+00
n_estimators	4.500e+02
max_depth	1.800e+01
learning_rate	1.000e-01
gamma	4.000e-01
colsample_bytree	4.000e-01

Kết quả tốt nhất

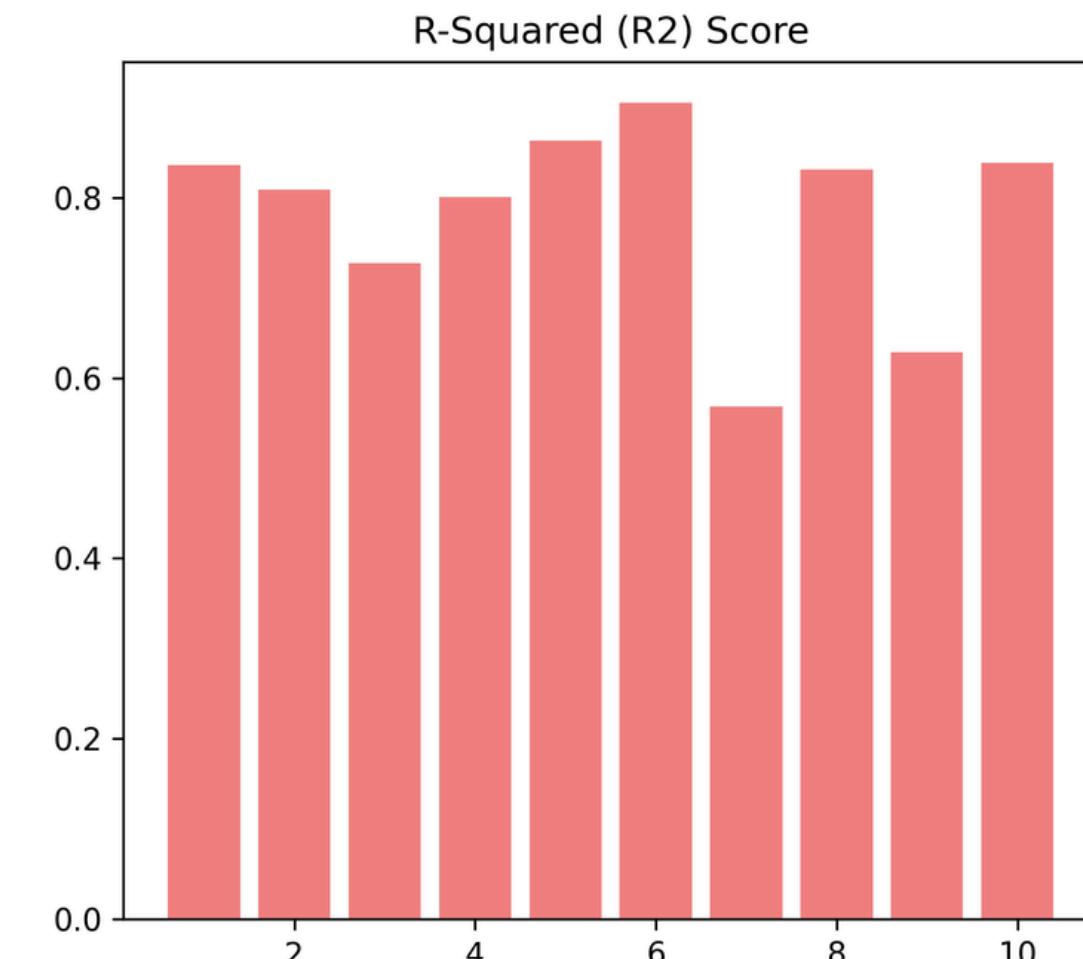
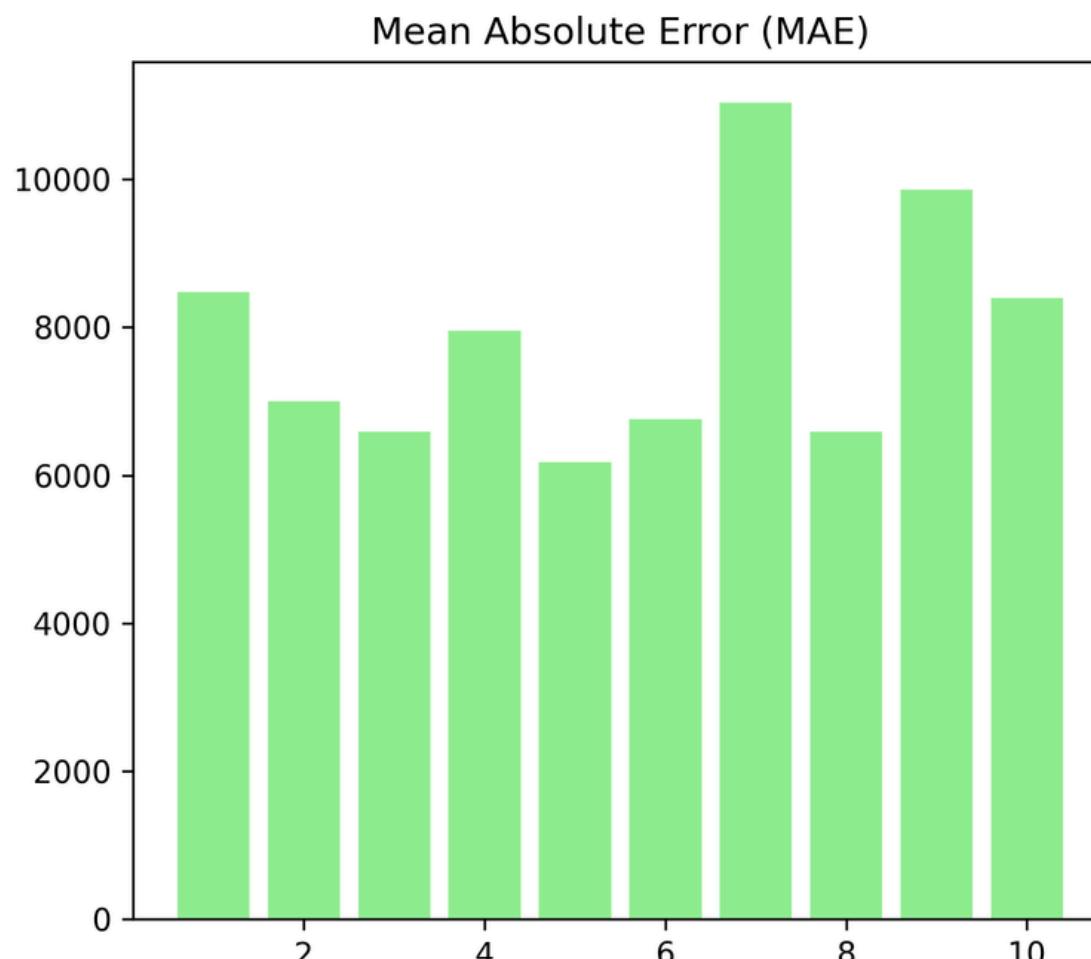
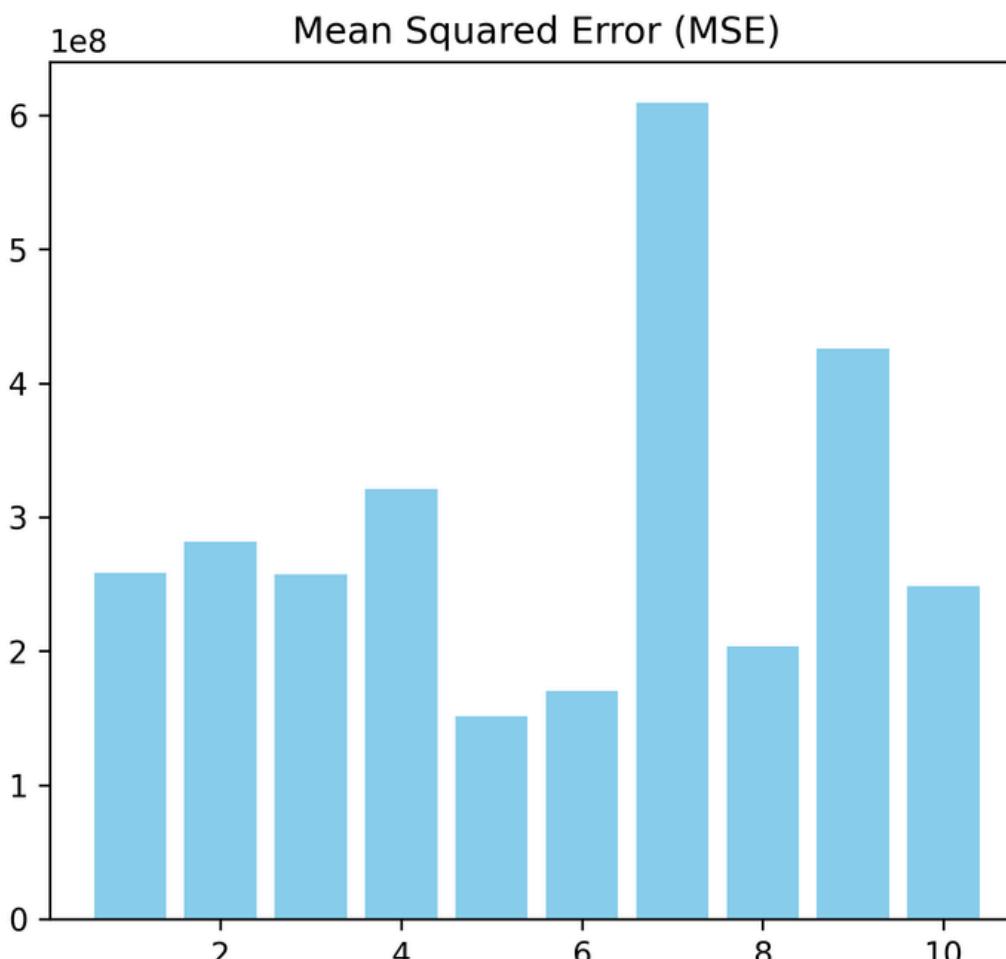
Mean Squared Error (MSE): 248466344.827

Mean Absolute Error (MAE): 7420.090

R-squared (R2) Score: 0.842

Data Modeling - Evaluation

Giá trị của các độ đo đánh giá trên từng Fold



Mean MSEs: 307139547.628

Mean MAEs: 8122.805

Mean R2s: 0.772

6. KẾT LUẬN

Các biến được sử dụng & Mô hình tốt nhất

Các biến được sử dụng: Company Name, job_simplified, seniority, Headquarters, Location, job_state, Industry, Type of ownership, Sector, Revenue, Size, Rating Category, Python_yn, Spark, AWS_yn, Average Salary

Mô hình tốt nhất: XGBoost

Các tham số tốt nhất

Điểm số tốt nhất: 0.727

Các tham số tốt nhất:

	reg_lambda	reg_alpha	n_estimators	max_depth	learning_rate	gamma	colsample_bytree
Value	1.000e-05	1.000e+00	4.500e+02	1.800e+01	1.000e-01	4.000e-01	4.000e-01

Kết quả tốt nhất

Mean Squared Error (MSE): 248466344.827

Mean Absolute Error (MAE): 7420.090

R-squared (R2) Score: 0.842

Cảm ơn

Mọi người đã lắng nghe!

Nhóm 19,
Khoa KHMT

