جامعة الملك سعود
King Saud University
كلية علوم الحاسب والمعلومات
قسم هندسة البرمجيات

LINKME

SWE 485

Presented By :

Leena Alsalman
Lamyaa Abukhalil
Jumanah Alwakeel
Layan Alhababi
Nouf Almandil

Fall 2025

Problem statement

Nowadays, many people are confused about which skills they need to acquire to get jobs they aspire to due to the rapid evolution of technologies and industries. New tools , programming languages , business practices and many other innovations are introduced regularly making the decision of which skills are really needed to acquire even more difficult.

This Project Aims to solve this Problem

1. Classify job categories
2. identify the skills people need to acquire for each job
3. Understand the market demands in a better way

Business value

- Data Driven Guidance for careers
- Scalability and automation
- Enhanced job market understandability
- better Upcoming skills decsisions

Value Proposition

A data-driven approach to employability and recruitment that transforms large volumes of LinkedIn job postings into clear, actionable, and highly useful skill recommendations.

Data Sources

LinkedIn job postings dataset from HuggingFace (the dataset)

Key Features Used

Job Title, Job Countryand Required Skills

These were the primary fields used for classification, clustering, and skill analysis.

Collection Methods

The dataset was collected from LinkedIn job postings using web scraping. It was provided in a cleaned CSV format.

- EDA

We counted the most common job titles, analyzed how these titles are distributed across countries, and detected missing values to ensure data quality before further processing.

- Data Cleaning & Preprocessing

We removed any rows with missing values and normalized the text for Job Title, Country, and Skills to ensure consistent formatting. To reduce overlap and prevent unstable predictions, we limited the skill set to the Top 30 most frequent skills. Additionally, we sampled 50 jobs per title for each country to create a balanced dataset and support fair, reliable model training.

- **Supervised Learning**

Build a multi label classifier. Learns from Job Title and Country, then predicts top skills using Logistic Regression and SVM, and then evaluates and compares both models. SVM performed best overall.

- **Unsupervised Learning**

Clustered jobs using Job Title and Country, selected the optimal number of clusters using the Elbow and Silhouette methods, applied a K-Means clustering model, and evaluated the cluster quality. Then combined clusters with TF IDF text features to train & evaluate LogReg and SVM for skill prediction.

- **Generative AI Integration**

Used a Generative AI model to transform each job's title, country, cluster, and skills into personalized career guidance and tailored skill-learning recommendations.

جامعة الملك سعود
King Saud University
كلية علوم الحاسب والمعلومات
قسم هندسة البرمجيات

Languages & Libraries

- Python
- Pandas
- matplotlib
- NumPy
- Scikit-learn
- ast

Platform

- Google Colab
- Jupyter Notebooks
- VS code
- Git hub

جامعة الملك سعود
King Saud University
كلية علوم الحاسب والمعلومات
قسم هندسة البرمجيات

KEY INSIGHTS

Our journey began with a simple question: **What is the job market really asking for?**

Using a LinkedIn dataset from Kaggle, we explored the landscape not to recommend jobs to individuals, but to understand market-wide trends and emerging skills.

1. The Market Is Data-Driven

The most common roles are Data Analyst, Data Engineer, and Data Scientist, showing a strong industry shift toward data-centric careers.

2. Companies Hide Salary Information

Missing value analysis revealed that salary fields are mostly empty, highlighting low transparency in job postings.

3. The U.S. Dominates Job Demand

Across top roles, the United States has the highest number of openings, indicating where global hiring is concentrated.

4. Skills Can Be Predicted

Supervised models showed that SVM performs best, accurately predicting key in-demand skills such as Python, SQL, and Spark.

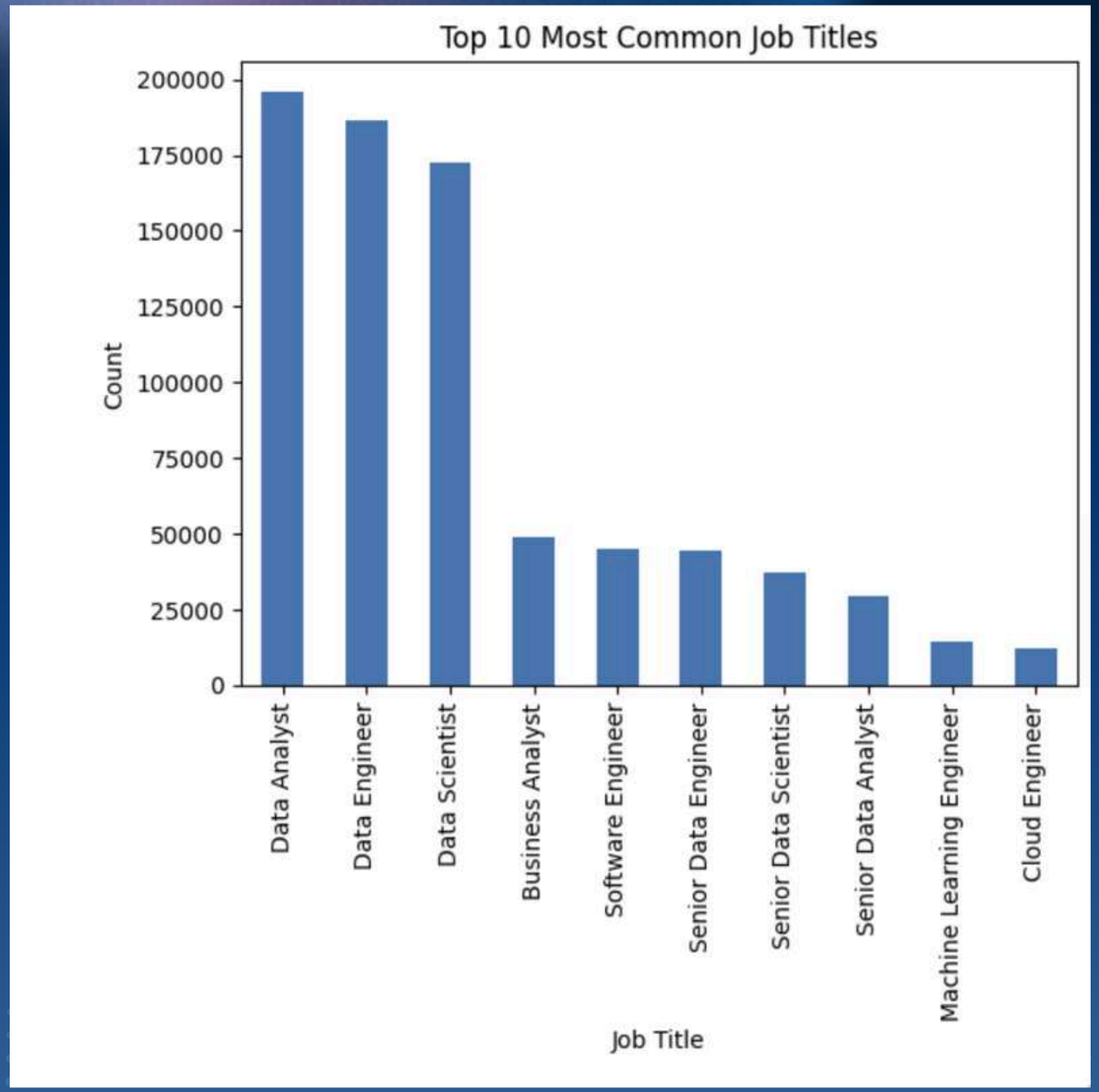
5. The Market Has Natural Job Clusters

Unsupervised learning revealed six clear job clusters, showing that roles naturally group by shared skill patterns.

6. Generative AI Makes Trends Actionable

Template 2 produced more contextual and detailed insights, helping summarize emerging skills for each cluster.

MOST COMMON JOB ROLES

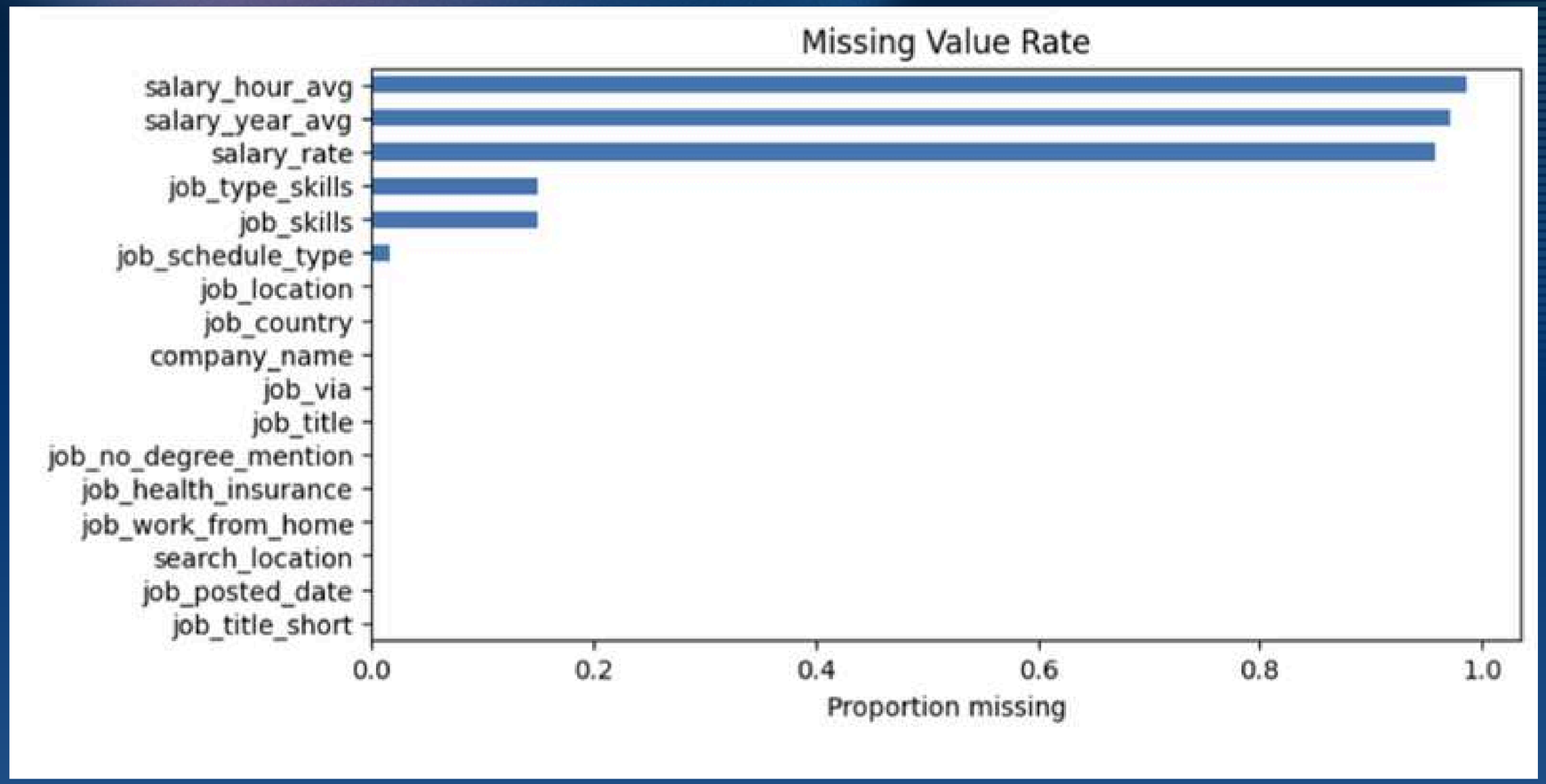


X-axis: Job title categories

Y-axis: Number of job postings

This graph represents the top 10 most common job titles in the dataset, and it shows that the top three roles are Data Analyst, Data Engineer, and Data Scientis

MISSING VALUE RATE PER FEATURE

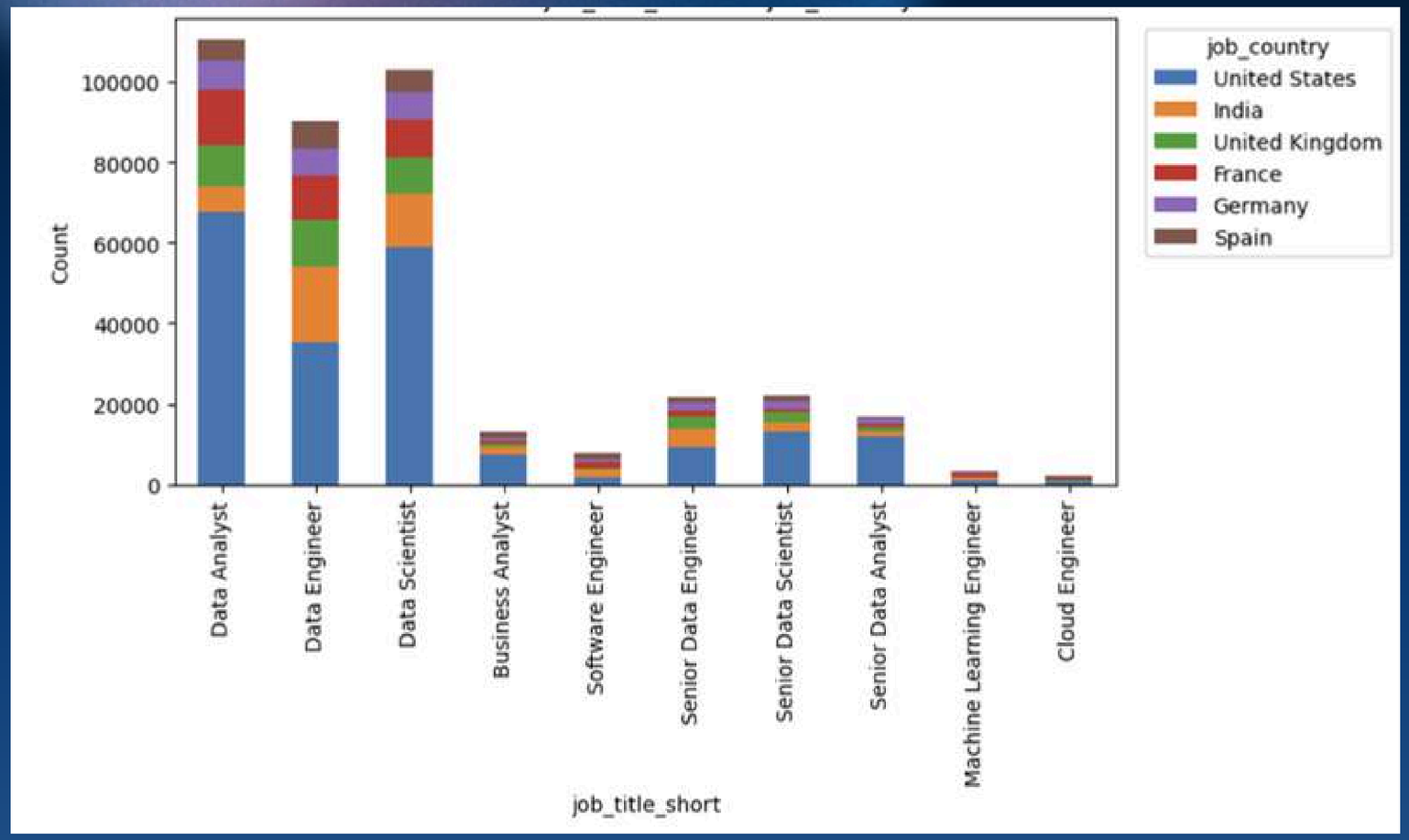


Y-axis: Dataset feature names

X-axis: Percentage of missing values (%)

This graph represents the percentage of missing values for each feature in the dataset, and it shows that the top three fields with the highest missing values are salary_hour_avg, salary_year_avg, and salary_rate

JOB TITLE DISTRIBUTION BY COUNTRY



X-axis: Job Title

Y-axis: Number of job postings

Colors: Countries

This graph represents the distribution of the top 10 job titles across different countries, and it shows that the United States has the highest number of postings overall.

SUPERVISED

SVM (Best Model)

- Highest F1-scores
- Better accuracy
- Works well with short job titles
- More balanced skill predictions

Logistic Regression

- Lower F1-scores
- Lower accuracy
- Over-predicts skills
- Less balanced results

Model	Accuracy	Micro F1-score	Macro F1-score	Precision	Recall
SVM	0.137296	0.561050	0.442640	0.479298	0.667237
Logistic Regression	0.115379	0.439617	0.386916	0.435792	0.736208

Why did SVM perform better?

- Job titles are short text, and SVM works better for short text classification.
- It separates decision boundaries between multiple skills more effectively.
- It handles imbalanced skills better.
- Logistic Regression had very high recall for some skills but much lower precision, often predicting too many skills per job.

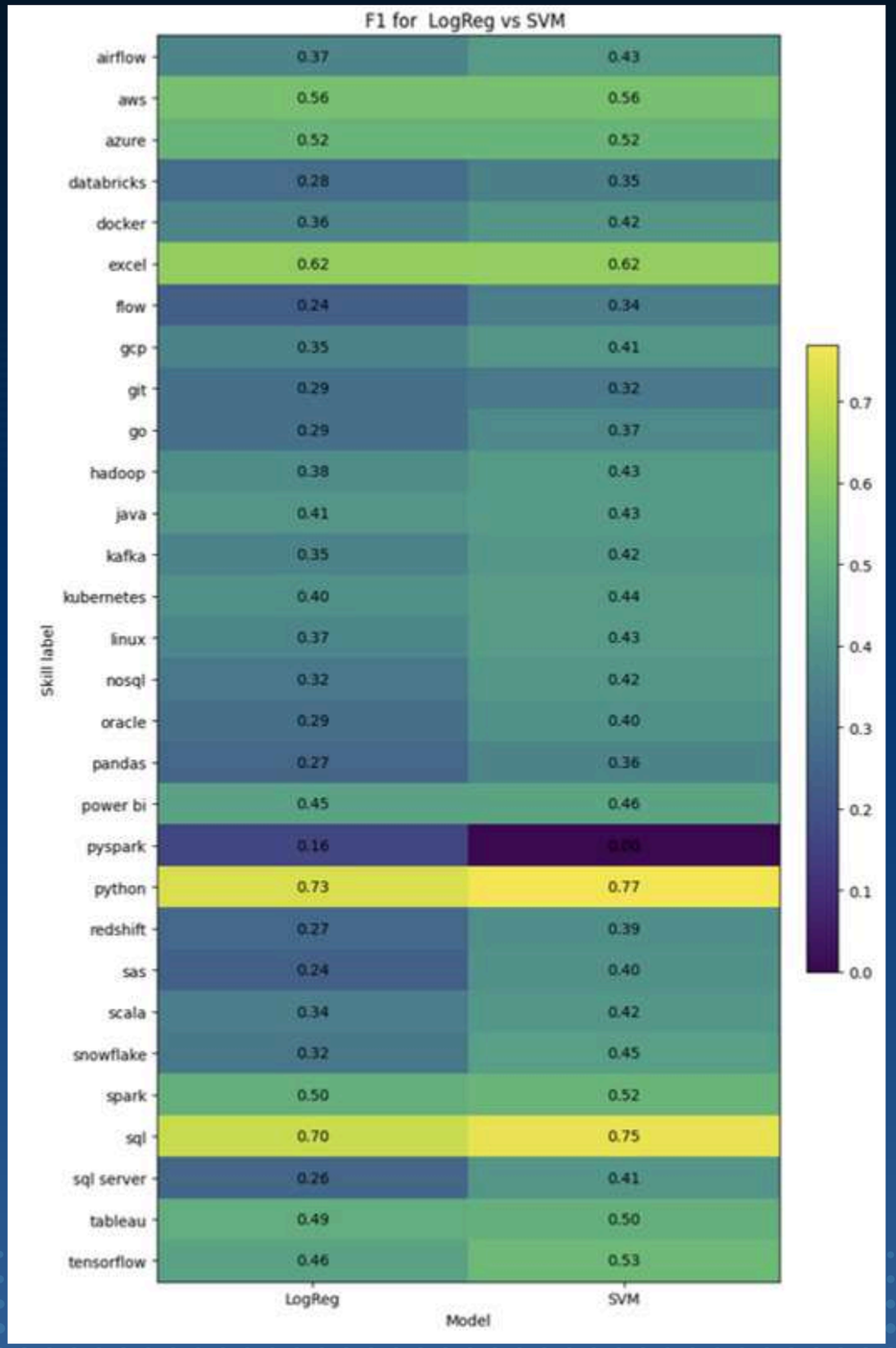
Observation

Common skills are predicted best
Rare skills are harder
Country improves predictions

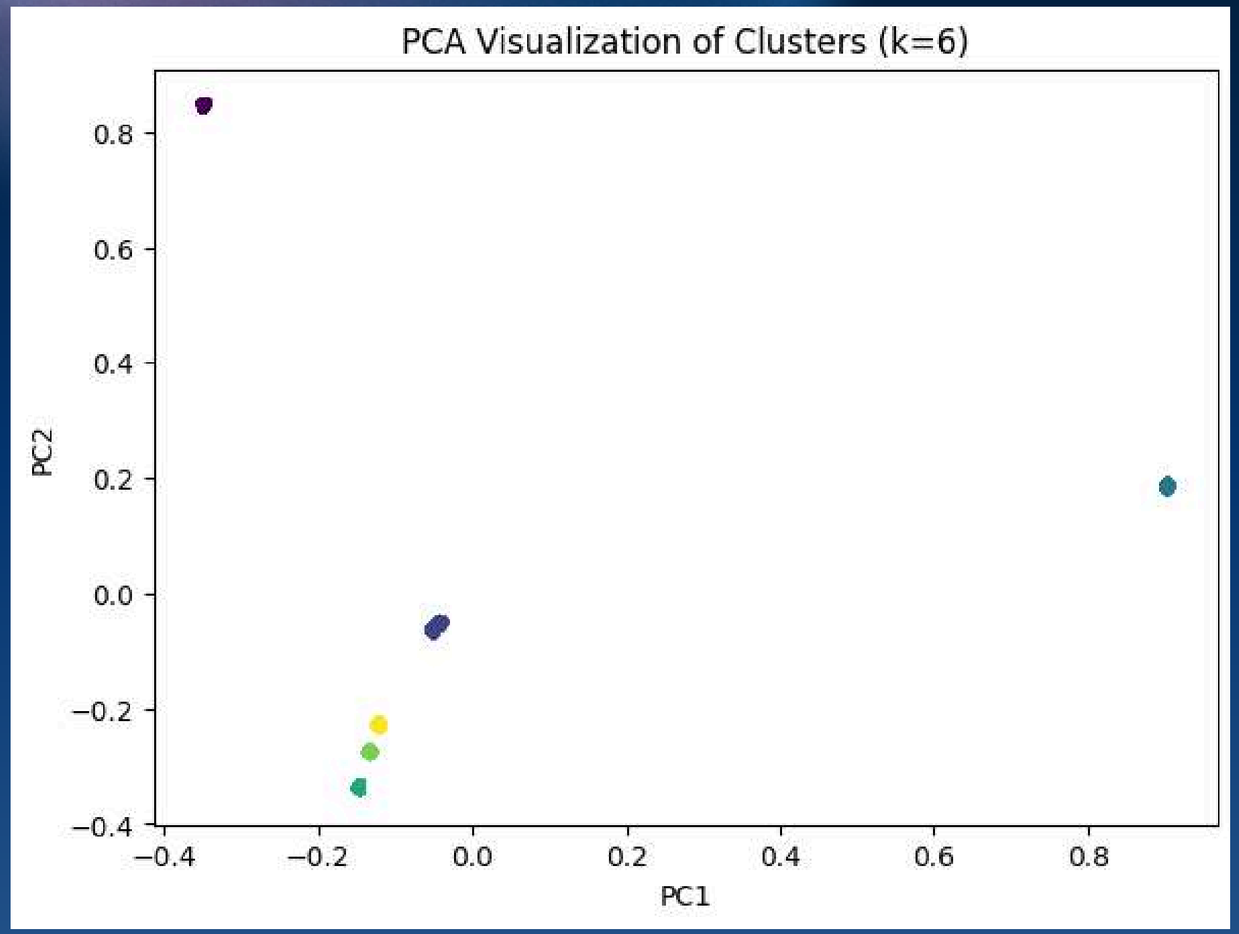
SUPERVISED

This heatmap represents the F1-score for each skill label using Logistic Regression and SVM, and it shows that SVM achieves higher F1-scores for most skills, especially Python, SQL, Spark, and TensorFlow.

Y-axis: Skill labels
X-axis: Models (Logistic Regression, SVM)
Color scale: F1-score (0.0 to 1.0)



UNSUPERVISED



This plot shows how our job data groups into six clusters after reducing it to two dimensions using PCA. Each point is a job posting, and each color represents a different cluster. The separation between the colored groups means the algorithm successfully found distinct job groups in the data.

جامعة الملك سعود
King Saud University
كلية علوم الحاسب والمعلومات
قسم هندسة البرمجيات

UNSUPERVISED

This table displays a sample of job postings after clustering, showing each job's title, country, its extracted skills, and the cluster it was assigned to. It helps confirm that similar roles—such as multiple Business Analyst postings from the same country—were grouped into the same cluster, indicating that the clustering model is correctly identifying and grouping related job types.

job_title_short	job_country	cluster	cluster
business analyst	afghanistan	['go', 'oracle']	5
business analyst	afghanistan	sas	5
business analyst	afghanistan	['go', 'oracle']	5
business analyst	afghanistan	['go', 'oracle']	5
business analyst	afghanistan	['sas']	5

جامعة الملك سعود
King Saud University

كلية علوم الحاسب والمعلومات
قسم هندسة البرمجيات

GENERATIVE AI

What This Table Shows

- We randomly selected 3 job postings from our clustered dataset.
- For each job, we passed its job title, country, skills, and cluster into two different AI prompt templates.
- The table shows the outputs generated by each template side by side.

job_title_short	job_country	cluster	skills_used	template_1_output	template_2_output
software engineer	liberia	4	[python,was]	For a software engineer role in Liberia, Pytho...	I'm thrilled to help you, and I'm excited to s...
software engineer	tanzania	4	[python, go, azure, databricks]	Based on the provided information, here's the ...	I'm excited to help you succeed as a software ..
cloud engineer	peru	1	[python, sql, pandas, flow]	Based on the role of a cloud engineer, here's ...	**Welcome, Cluster 1 User from Peru!**\n\nI'm ...

What We Observed

- Template 1 → Short, simple, and informative.
- Template 2 → More personalized, detailed, and motivational.

Template 2 uses the country, cluster, and skills more effectively.

جامعة الملك سعود
King Saud University
كلية علوم الحاسب والمعلومات
قسم هندسة البرمجيات

Summary

Our analysis showed that data-focused roles dominate the job market, the U.S. leads global hiring, and skills like Python and SQL can be accurately predicted using ML making them critical for career readiness.

Recommendations

- Job seekers should prioritize core data skills (Python, SQL, Spark, cloud).
- Educators should align programs with market-demand skills.
- Platforms can use our models to offer personalized skill guidanc

Future work

Enhance predictions with deeper models, expand datasets to more regions, and develop a full AI-powered career guidance platform.

We learned how important data cleaning is in real-world datasets, how model choice impacts multi-label predictions, and how clustering reveals natural job patterns. We also saw the value of Generative AI in turning analytics into clear, personalized insights.