Data Analysis

Comprehensive Data Cleaning & Exploratory Analysis of Job Market Trends

team3

2025-04-29

|  |
| --- |
| Note |
| This is the final integrated report combining all sections: data cleaning, skill gap analysis, machine learning methods, and final insights. |

title: “Salary and Compensation Trends in AI Careers” author: “Team 3” date: today format: html: toc: true toc-depth: 2 toc-exclude: [“Welcome to Our Research”] bibliography: references.bib csl: csl/econometrica.csl

# Welcome to Our Research

## Research Rationale

Our team chose **Salary Disparities Across Disciplines: Who Benefits Most from AI’s Rise?**  as our research topic. Due to the rapid development of AI chatbots and substantial infrastructure investments by companies such as OpenAI, Google, NVIDIA, and DeepSeek, a significant restructuring of the current workforce appears inevitable. Some existing occupations will diminish or potentially disappear, while new roles will emerge and command higher compensation. Understanding these changes, identifying potential trends, and analyzing reliable data regarding AI’s impact are of paramount importance.

## Why is this topic important?

The rapid advancement of AI is transforming the workforce by reshaping job opportunities and salary structures. While AI-driven technologies enhance productivity and facilitate the emergence of new roles, they also contribute to job displacement, particularly in routine-based occupations.

Empirical studies have demonstrated that automation has been a key driver of income inequality (Adhikari (2024)), disproportionately impacting middle-skilled and less-educated workers. Conversely, AI-driven professions are experiencing new work opportunities, particularly in technology-related fields, necessitating a shift in workforce skills and adaptability.

The adoption of AI technologies brings **wage volatility and job opportunities** across different income groups. While AI roles often require unique, high-demand skillsets within specific occupations, it is overall important to implement policy measures to reduce the variations in industry adoption and workforce accessibility. As AI reshapes job requirements, reskilling becomes imperative, particularly for non-technical workers who are at higher risk of displacement. Addressing the digital skills gap through targeted training programs is essential to ensuring workforce adaptability and mitigating the socioeconomic disruptions caused by AI-driven labor market transformations.

## Trend Analysis of AI Development and Salary Fluctuations

The rapid advancement of AI has led to an increasing market demand for AI-related skills, significantly influencing employment trends and salary structures in the United States. Studies suggest that various job sectors are highly susceptible to AI-driven transformation, including those in traditionally high-skilled industries (Colombo et al. (2024)).

Similarly, in China, the demand for AI professionals surged in 2024, particularly in specialized domains such as **healthcare and applied sciences**, as companies sought to attract top talent in these fields. Salaries for AI specialists have risen substantially. Moreover, the growing competition in the AI sector has reshaped existing salary distributions. AI-related professions offer **higher salary premiums and employment benefits** compared to traditional IT engineering roles, potentially exacerbating the wage disparity between AI professionals and non-AI professionals, thereby influencing labor market equilibrium (Stone, Lukaszewski, and Johnson (2024)).

AI is fundamentally transforming labor markets across various industries, exerting a significant influence on **wage structures depending on job characteristics**. Repetitive rule-based tasks, such as **basic data entry or customer service automation**, are highly susceptible to AI-driven automation, reducing the demand for such roles in the labor market, thereby cutting down employment opportunities and lowering compensation obtained. For instance, AI has been widely adopted in **human resources management candidate screening**, prompting some employers to lessen workforce hiring needs (Sezgin (2023)).

Conversely, occupations requiring **advanced cognitive skills, creativity, or interpersonal communication**—such as those in **professional services**—are less likely to be fully replaced by AI and may instead benefit from hybrid automation models.

* **Medical field**: AI-powered diagnostic tools have enhanced precision and efficiency in medical decision-making, but final assessments still rely heavily on human expertise (Ansari and Ansari (2024)).
* **Legal sector**: AI-driven automation has accelerated case analysis and legal procedures, but **complex legal judgments continue to rely on human practitioners**. Harvard Law Professor **David Wilkins** has found that while generative AI has the potential to transform legal practices, the primary role remains that of a **supportive tool rather than a replacement**.
* **Finance and Governance**: Industries with a focus on **financial and governance decision-making** may experience **wage polarization** due to AI-induced advancements, where professionals may experience **wage growth** due to AI-driven productivity enhancements.

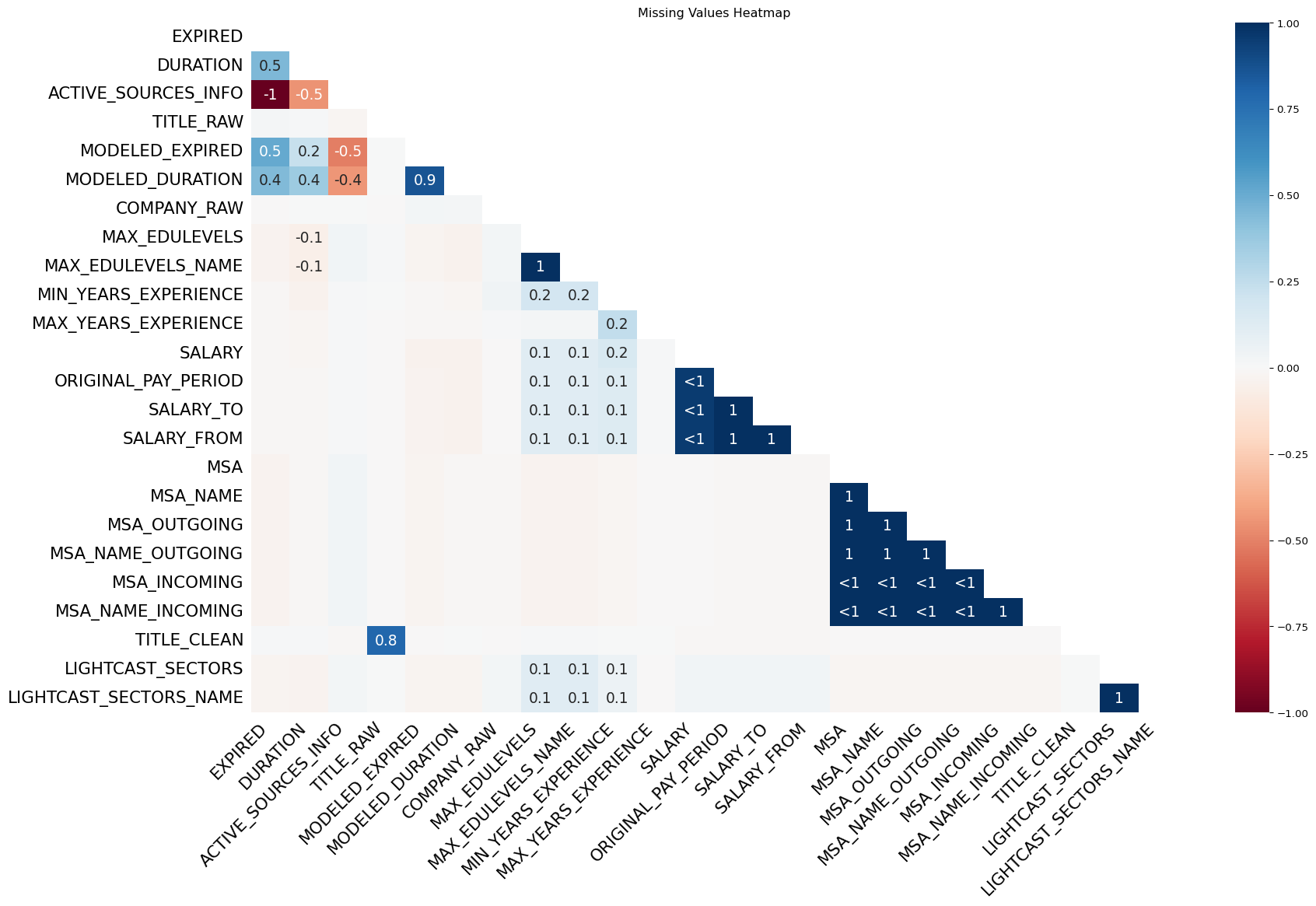
## What do you expect to find in your analysis?

Our research encompasses various aspects of **salary disparities**. Specifically, we aim to investigate: - **Income and job distribution differences**, including an analysis of which **geographical regions** see the highest demand and wage disparities for both AI-related and traditional professional roles. - **Comprehensive analysis of recent datasets**, identifying the most sought-after required skills among job listings. - **The differential impact of AI-driven changes on high-skill versus middle-skill occupations**.

## References

## Data Cleaning & Preprocessing

import pandas as pd  
import missingno as msno  
import matplotlib.pyplot as plt  
import plotly.express as px  
  
# 1. Read the CSV file  
df = pd.read\_csv("lightcast\_job\_postings.csv")  
  
# 2. Columns to drop (if they exist)  
columns\_to\_drop = [  
 "ID", "URL", "ACTIVE\_URLS", "DUPLICATES", "LAST\_UPDATED\_TIMESTAMP",  
 "NAICS2", "NAICS3", "NAICS4", "NAICS5", "NAICS6",  
 "SOC\_2", "SOC\_3", "SOC\_5"  
]  
# Filter out columns that actually exist in the DataFrame to avoid KeyError  
valid\_cols\_to\_drop = [col for col in columns\_to\_drop if col in df.columns]  
df.drop(columns=valid\_cols\_to\_drop, inplace=True)  
  
# 3. Visualize missing values  
msno.heatmap(df)  
plt.title("Missing Values Heatmap")  
plt.show()  
  
# 4. Basic missing value processing  
df.dropna(thresh=len(df) \* 0.5, axis=1, inplace=True) # If a column has more than 50% missing values, delete it.  
  
# If there is a 'SALARY' column, fill missing values with the median.  
if "SALARY" in df.columns:  
 df["SALARY"].fillna(df["SALARY"].median(), inplace=True)  
# If there is an 'Industry' column, fill missing values with 'Unknown'.  
if "Industry" in df.columns:  
 df["Industry"].fillna("Unknown", inplace=True)  
  
# 5. Remove duplicates  
df.drop\_duplicates(subset=["TITLE", "COMPANY", "LOCATION", "POSTED"], keep="first", inplace=True)



### Job Postings by Industry

if "Industry" in df.columns:  
 fig = px.bar(df["Industry"].value\_counts(), title="Job Postings by Industry")  
 fig.show()

**Explanation:**  
This bar chart shows the number of job postings across different industries. It highlights which sectors are most active in recruiting, allowing us to identify high-demand areas such as technology, healthcare, or finance.

### Salary Distribution by Industry

if "SALARY" in df.columns and "Industry" in df.columns:  
 fig = px.box(df, x="Industry", y="SALARY", title="Salary Distribution by Industry")  
 fig.show()

**Explanation:**  
This box plot illustrates how salary ranges differ across industries. It helps reveal not only the median salary but also the variability and presence of outliers, giving insight into income inequality or competitive compensation within certain sectors.

### Remote vs. On-Site Jobs

if "REMOTE\_TYPE\_NAME" in df.columns:  
 fig = px.pie(df, names="REMOTE\_TYPE\_NAME", title="Remote vs. On-Site Jobs")  
 fig.show()

Unable to display output for mime type(s): text/html

Unable to display output for mime type(s): text/html

**Explanation:**  
This pie chart breaks down the proportion of jobs by remote type. It shows how many roles are fully remote, hybrid, or entirely on-site, offering insights into post-pandemic work trends and flexibility offered by employers.

### Job Postings Over Time

if "POSTED" in df.columns:  
 df['POSTED'] = pd.to\_datetime(df['POSTED'], errors='coerce')  
 postings\_over\_time = df['POSTED'].value\_counts().sort\_index()  
 fig = px.line(x=postings\_over\_time.index, y=postings\_over\_time.values, labels={'x': 'Date Posted', 'y': 'Number of Job Postings'}, title="Job Postings Over Time")  
 fig.show()

Unable to display output for mime type(s): text/html

**Explanation:**

This line graph tracks how job postings have changed over time. Peaks and troughs may indicate hiring cycles, economic events, or seasonal effects in the job market.

title: “Skill Gap Analysis”  
format: html kernel: python3  
execute: echo: true  
eval: true

## 3.1 Skill Gap Analysis

### 3.1.1 Team Skill Self-Assessment

import pandas as pd  
  
# Team members and their self-rated proficiency (1–5)  
# Team members and their self-rated proficiency (1–5)  
skills\_data = {  
 "Name": ["Deyang", "Yani", "Jiapei", "Junhao"],  
 "Python": [4, 3, 4, 3],  
 "SQL": [3, 3, 4, 3],  
 "Machine Learning": [2, 5, 5, 2],  
 "Cloud Computing": [1, 4, 2, 3],  
 "R": [3, 5, 4, 2], #   
 "AWS": [4, 4, 2, 3], #   
 "Git": [4, 3, 2, 1], #   
 "Excel": [3, 4, 5, 2], #   
}  
  
df\_skills = pd.DataFrame(skills\_data).set\_index("Name")  
df\_skills

|  | Python | SQL | Machine Learning | Cloud Computing | R | AWS | Git | Excel |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Name |  |  |  |  |  |  |  |  |
| Deyang | 4 | 3 | 2 | 1 | 3 | 4 | 4 | 3 |
| Yani | 3 | 3 | 5 | 4 | 5 | 4 | 3 | 4 |
| Jiapei | 4 | 4 | 5 | 2 | 4 | 2 | 2 | 5 |
| Junhao | 3 | 3 | 2 | 3 | 2 | 3 | 1 | 2 |

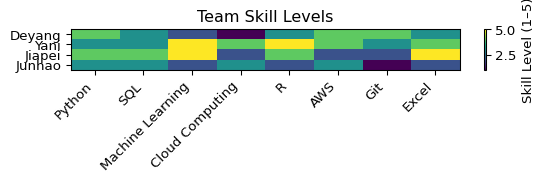
3.1.2 Compare to Market Demand

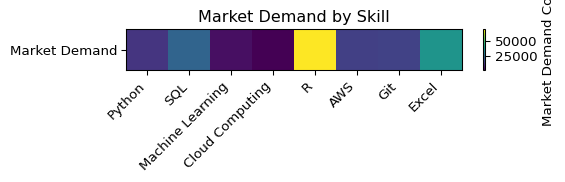
# Load raw job postings data  
raw\_df = pd.read\_csv("lightcast\_job\_postings.csv")  
  
# Drop irrelevant columns  
columns\_to\_drop = [  
 "ID", "URL", "ACTIVE\_URLS", "DUPLICATES", "LAST\_UPDATED\_TIMESTAMP",  
 "NAICS2", "NAICS3", "NAICS4", "NAICS5", "NAICS6",  
 "SOC\_2", "SOC\_3", "SOC\_5"  
]  
valid\_cols\_to\_drop = [col for col in columns\_to\_drop if col in raw\_df.columns]  
raw\_df.drop(columns=valid\_cols\_to\_drop, inplace=True)  
  
# Drop columns with >50% missing  
raw\_df.dropna(thresh=len(raw\_df) \* 0.5, axis=1, inplace=True)  
  
# Fill missing values  
if "SALARY" in raw\_df.columns:  
 raw\_df["SALARY"].fillna(raw\_df["SALARY"].median(), inplace=True)  
if "Industry" in raw\_df.columns:  
 raw\_df["Industry"].fillna("Unknown", inplace=True)  
  
# Remove duplicates  
raw\_df.drop\_duplicates(subset=["TITLE", "COMPANY", "LOCATION", "POSTED"], keep="first", inplace=True)  
  
# Count keyword occurrences  
top\_skills = df\_skills.columns.tolist()  
job\_text = raw\_df["BODY"].fillna("")  
skill\_counts = {s: job\_text.str.contains(s, case=False).sum() for s in top\_skills}  
  
# Append demand row  
df\_skills.loc["Market Demand"] = [skill\_counts[s] for s in top\_skills]  
df\_skills

|  | Python | SQL | Machine Learning | Cloud Computing | R | AWS | Git | Excel |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Name |  |  |  |  |  |  |  |  |
| Deyang | 4 | 3 | 2 | 1 | 3 | 4 | 4 | 3 |
| Yani | 3 | 3 | 5 | 4 | 5 | 4 | 3 | 4 |
| Jiapei | 4 | 4 | 5 | 2 | 4 | 2 | 2 | 5 |
| Junhao | 3 | 3 | 2 | 3 | 2 | 3 | 1 | 2 |
| Market Demand | 11784 | 23207 | 3974 | 1303 | 69181 | 14245 | 14696 | 36384 |

3.1.3 Visualize Skill Gaps

import matplotlib.pyplot as plt  
  
# Team skill levels heatmap  
plt.figure(figsize=(6,2))  
plt.imshow(df\_skills.iloc[:-1], aspect="auto")  
plt.colorbar(label="Skill Level (1–5)")  
plt.yticks(range(len(df\_skills.index)-1), df\_skills.index[:-1])  
plt.xticks(range(len(df\_skills.columns)), df\_skills.columns, rotation=45, ha="right")  
plt.title("Team Skill Levels")  
plt.tight\_layout()  
plt.show()  
  
# Market demand heatmap  
plt.figure(figsize=(6,2))  
plt.imshow([df\_skills.loc["Market Demand"]], aspect="auto")  
plt.colorbar(label="Market Demand Count")  
plt.yticks([0], ["Market Demand"])  
plt.xticks(range(len(df\_skills.columns)), df\_skills.columns, rotation=45, ha="right")  
plt.title("Market Demand by Skill")  
plt.tight\_layout()  
plt.show()





hj

Based on the heatmap comparison of each person’s self-ratings (1–5) against the normalized market-demand scores, here’s a concise, paragraph-style improvement plan:

Deyang shows the largest gaps in Cloud Computing and R. To close these, start with an AWS Fundamentals micro-course (Coursera or AWS’s own training) and follow up by building a small R Shiny dashboard using the free “R for Data Science” online book and the swirl R package for hands-on exercises. Aim to spend 2–3 hours per week on labs, then peer-review each other’s code in GitHub.

Yani needs to boost Machine Learning and AWS skills. I recommend Andrew Ng’s Machine Learning specialization on Coursera, paired with the AWS Certified Cloud Practitioner path on AWS Training. After completing each module, apply what you’ve learned by deploying a simple classification model on AWS Sagemaker and sharing the workflow in a team repo, so everyone can give feedback.

Jiapei would benefit most from deeper Cloud Computing practice and reinforcing R. Enroll in Google Cloud’s “Data Engineering” Qwiklabs quests and run through interactive R exercises via swirl. Host a 30-minute “teach-back” session after completing each mini-project—this both cements your own understanding and helps teammates pick up new tricks.

Junhao has room to grow in Excel and SQL. Take an “Excel Essentials” short course on LinkedIn Learning, then tackle Mode’s SQL tutorial problems. Organize weekly problem-solving sessions where one member presents a real-world dataset and the rest write SQL queries together. This combination of structured learning and peer collaboration will efficiently close the remaining gaps.

# Machine Learning Methods

## Overview

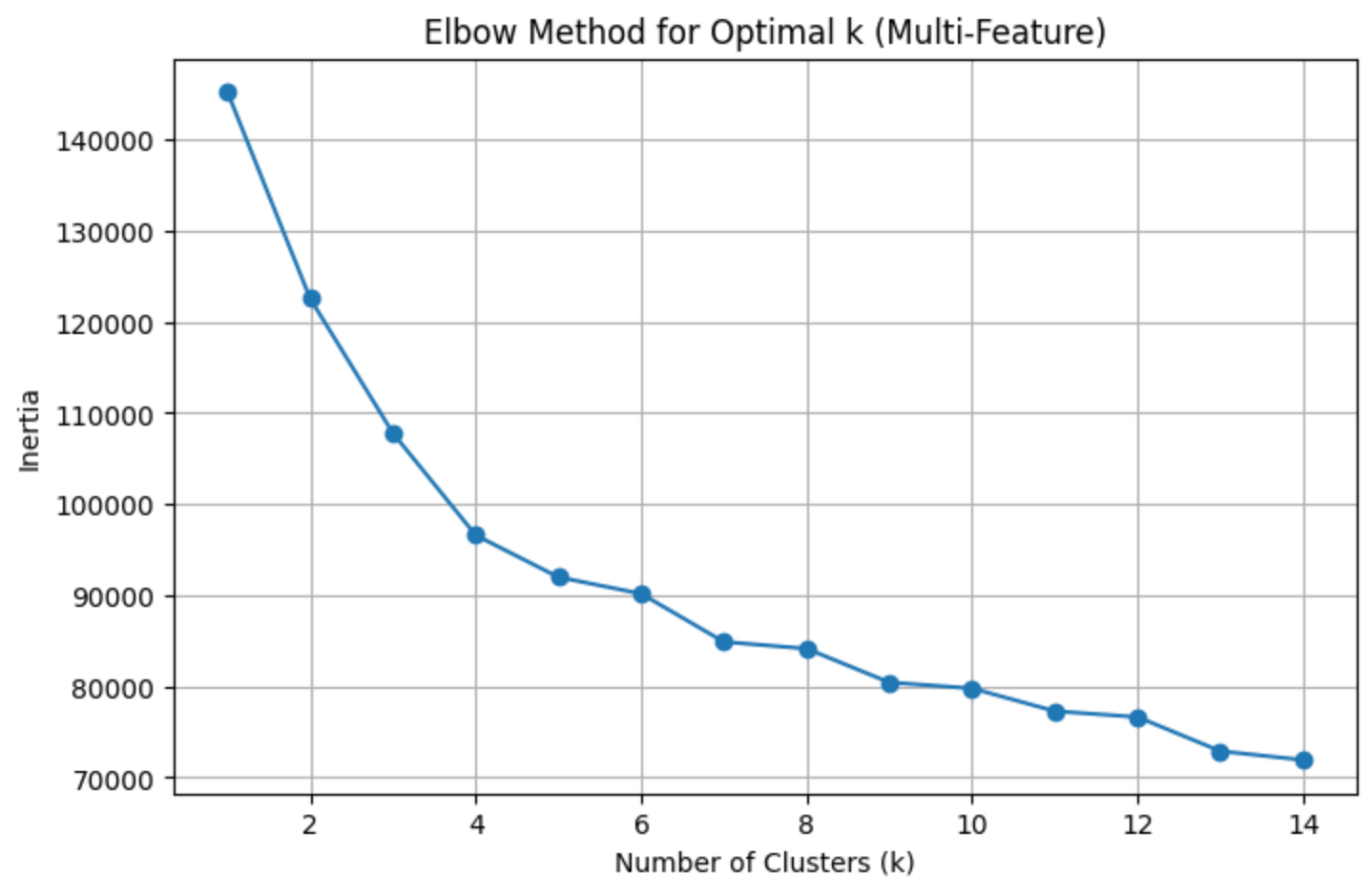
In this section, we applied two machine learning approaches to our job market dataset:  
- KMeans Clustering to segment job postings based on work type, industry, and education  
- Linear Regression to predict salary levels and assess the importance of different job features

## KMeans Clustering

We used the following features for clustering:

- NAICS\_2022\_5\_NAME (Industry)  
- EMPLOYMENT\_TYPE\_NAME (Full-time/Part-time)  
- REMOTE\_TYPE\_NAME (Work type)  
- EDUCATION\_LEVELS\_NAME (Minimum required education)

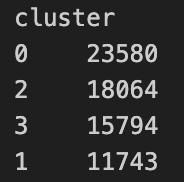
After one-hot encoding, we used the Elbow Method to select k=4.



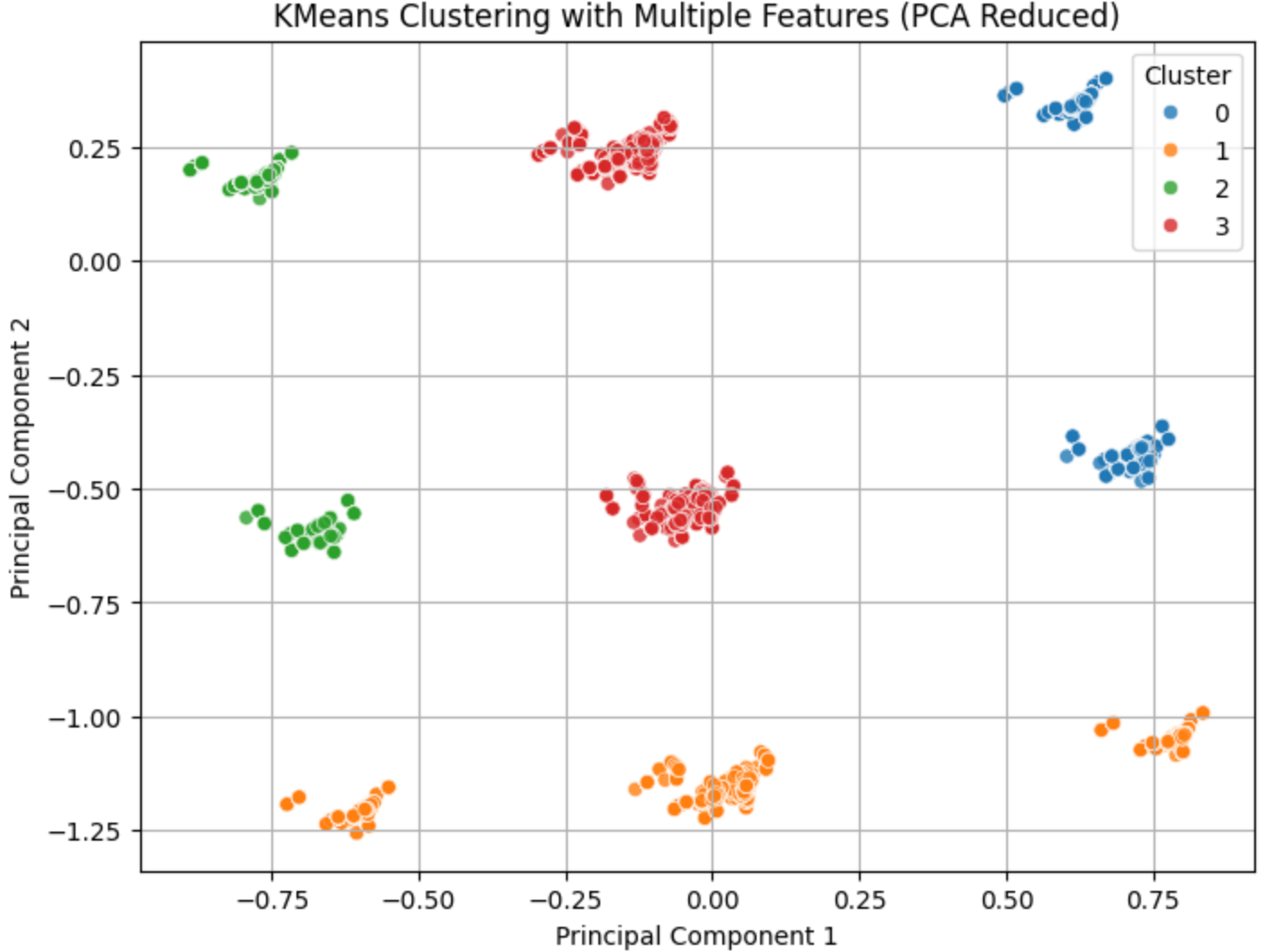
KMeans clustering code:

kmeans = KMeans(n\_clusters=4, random\_state=42)  
df\_kmeans['cluster'] = kmeans.fit\_predict(X)

### Cluster Profile Summary



### Interpretation



- Cluster 1: Most remote-friendly jobs.  
- Cluster 2 & 3: Tech jobs with high education requirements.  
- Cluster 0: Entry-level, mostly on-site jobs.

## Linear Regression for Salary Prediction

We used a linear regression model based on title, education, remote type, industry, and experience.

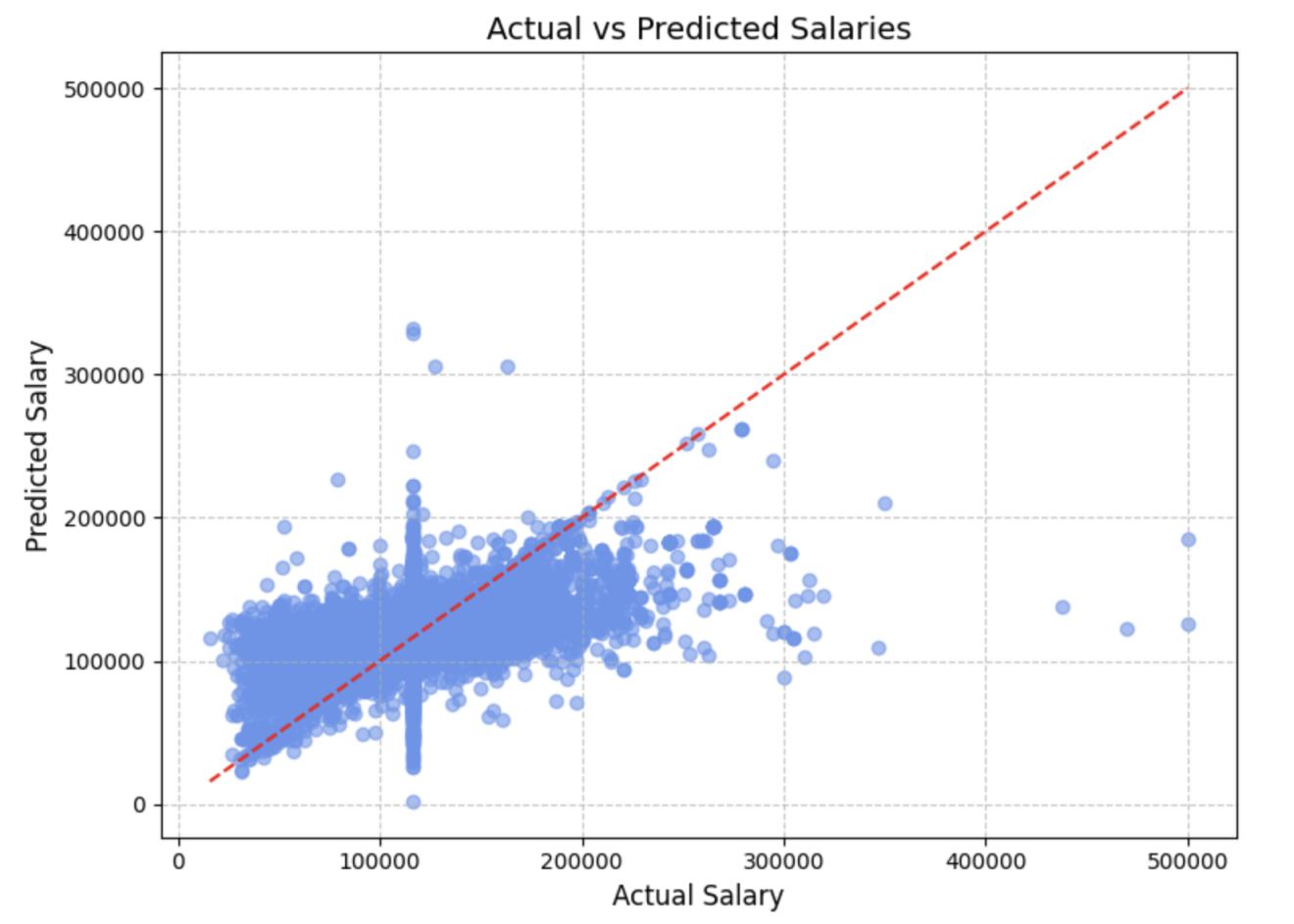
Model training code:

lr\_model = LinearRegression()  
lr\_model.fit(X\_train, y\_train)  
  
# Performance  
RMSE: 24,856  
R²: 0.291

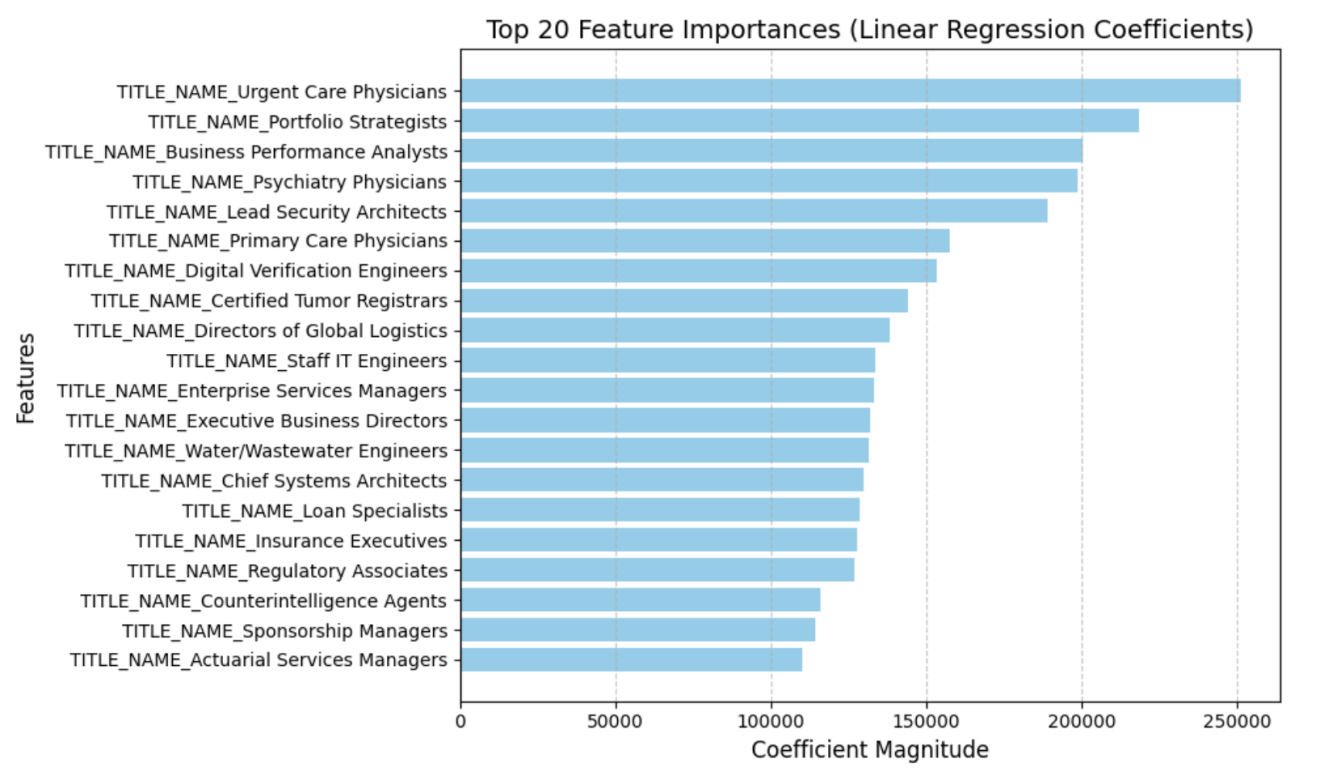
### Top Salary-Boosting Roles

- Urgent Care Physicians  
- Portfolio Strategists  
- Lead Security Architects  
- Digital Verification Engineers

### Model Performance



### Feature Importance



## Conclusion

These ML models help us:  
- Understand job clusters aligned with different work preferences  
- Identify job titles that drive salary differences  
- Provide strategic insights for job seekers

## Final analysis

In this group project, we started by conducting exploratory data analysis using Plotly. We visualized the job market from three perspectives: the number of job postings by industry, salary distributions, and the proportion of remote versus on-site jobs. These visualizations helped us gain a clearer understanding of the current job market landscape and trends, which in turn supported us in identifying more active industries and attractive roles—useful insights for making informed job application decisions.

Next, we focused on analyzing skill gaps by building a skill rating matrix for our team members. This allowed us to assess each person's proficiency across key technical areas. By comparing our skills with industry job requirements, we identified current shortcomings and proposed improvement plans. This process helped us better understand our levels in essential skills like Python, SQL, and Machine Learning, and guided us in creating targeted learning strategies. It also prepared us to better present our strengths—and honestly acknowledge our weaknesses—in future interviews.

Finally, we developed two modeling approaches: clustering and regression. We used KMeans clustering to categorize job types and applied multiple linear regression to predict salary levels based on variables like job title, location, and skills. These models not only supported deeper analysis but also strengthened our ability to apply statistical and machine learning techniques to real-world datasets. For roles such as data analyst, business intelligence, or HR tech, this kind of hands-on project experience is a valuable asset and can serve as a strong addition to our personal portfolios.

Adhikari, P. (2024): “[Exploring the nexus between artificial intelligence and job displacement: A literature review](https://jndmeerut.org/wp-content/uploads/2024/07/JND-1.pdf),” *Journal of National Development*, 37, 1–13.

Ansari, A., and A. Ansari. (2024): “[Consequences of AI induced job displacement](https://ijbat.com/index.php/IJBAT/article/view/18/31),” *International Journal of Business, Analytics, and Technology*, 2, 4–19.

Colombo, E., F. Mercorio, M. Mezzanzanica, and A. Serino. (2024): “[Towards the Terminator Economy: Assessing Job Exposure to AI Through LLMs](https://doi.org/10.48550/arXiv.2407.19204),”

Sezgin, E. (2023): “[Artificial intelligence in healthcare: Complementing, not replacing, doctors and healthcare providers](https://doi.org/10.1177/20552076231186520),” *Digital Health*, 9.

Stone, D. L., K. M. Lukaszewski, and R. D. Johnson. (2024): “[Will artificial intelligence radically change human resource management processes?](https://doi.org/10.1016/j.orgdyn.2024.101034)” *Organizational Dynamics*, 53, 101034.