Evaluating Personal Job Market Prospects in 2024

AD688 Final Report

Yixuan Yang Chengjie Lu Arohit Talari

April 29, 2025

Introduction

Background

In recent years, the job market has undergone significant transformation driven by technological advancements and changes in work dynamics. As graduate students preparing to enter a competitive job landscape, understanding these trends and aligning our skills accordingly is critical.

Project Goal

This project aims to assess personal job market readiness by analyzing industry trends, identifying skill gaps, and applying machine learning techniques to predict salary outcomes. Our ultimate objective is to propose personalized learning paths that enhance employability in the evolving market.

Methods Overview

The project combines data cleaning, exploratory analysis, machine learning modeling, and skill assessment. Publicly available job posting data was used to explore patterns, build predictive models, and benchmark team capabilities against market expectations.

Research Introduction

Research Questions

1. Which industries are generating the highest number of job postings in 2024?
2. How does salary distribution vary across industries and job types?
3. What is the current skill gap between our team and the market demands?
4. How can we use machine learning models to estimate job market value and identify important predictors?

Contribution

By linking data-driven market insights with personalized upskilling recommendations, this project provides both a strategic career roadmap and a framework for future job market analytics.

Data Analysis

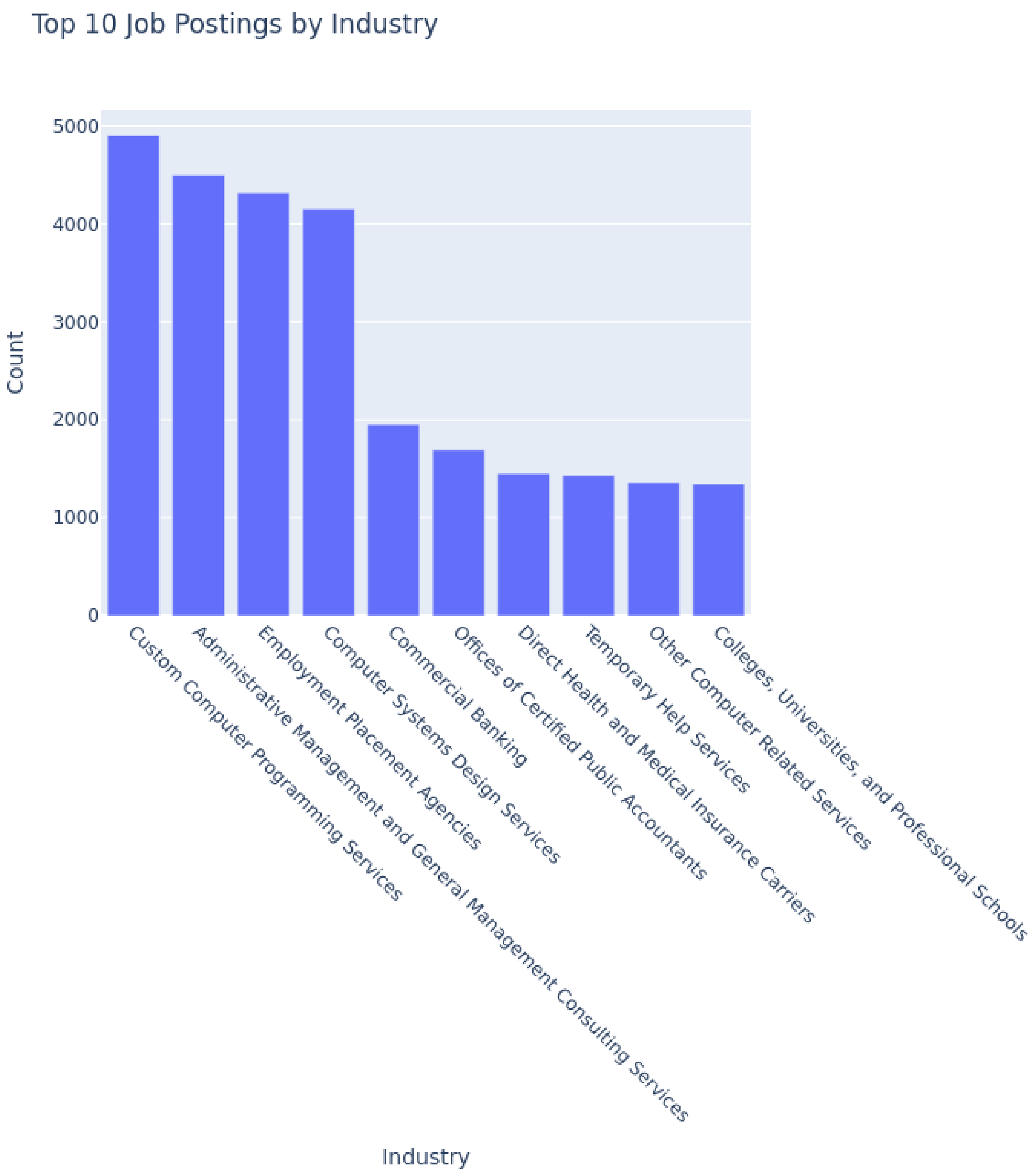
Comprehensive Data Cleaning & Exploratory Analysis of Job Market Trends

Yixuan Yang (Boston University)
Arohit Talari (Boston University)
Chengjie Lu (Boston University)

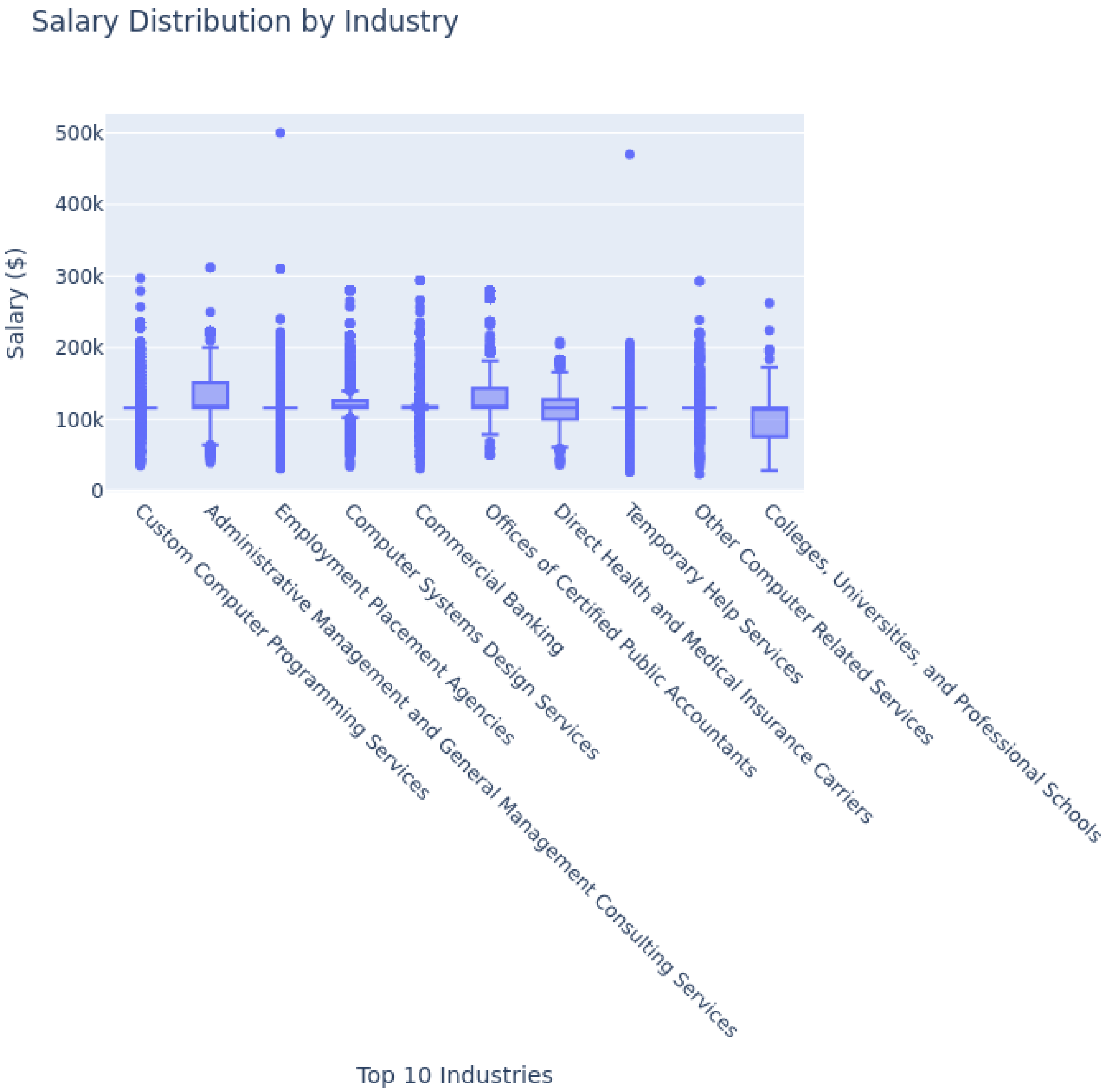
Data Preparation and Cleaning

```
LAST_UPDATED_DATE          0
POSTED                     0
EXPIRED                    0
DURATION                   0
SOURCE_TYPES               0
                          ..
LOT_V6_CAREER_AREA_NAME    0
LIGHTCAST_SECTORS          0
LIGHTCAST_SECTORS_NAME     0
NAICS_2022_6               0
NAICS_2022_6_NAME          0
Length: 99, dtype: int64
```

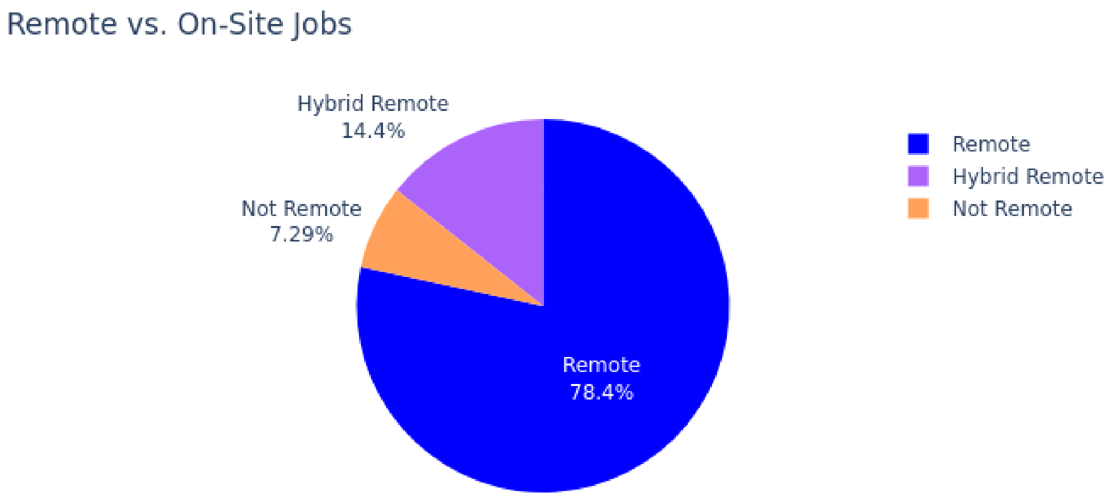
Data Visualization



The bar plot is used to display the top 10 highest number of job posting industries. The graph shows that computer related services are standing out, management services and employment placement agencies also have double the amount of job postings than others in this category.



The box plot presents the salary distribution across the top 10 industries with the highest number of job postings. By reducing the number of categories and adjusting the axis labels, we improve readability.

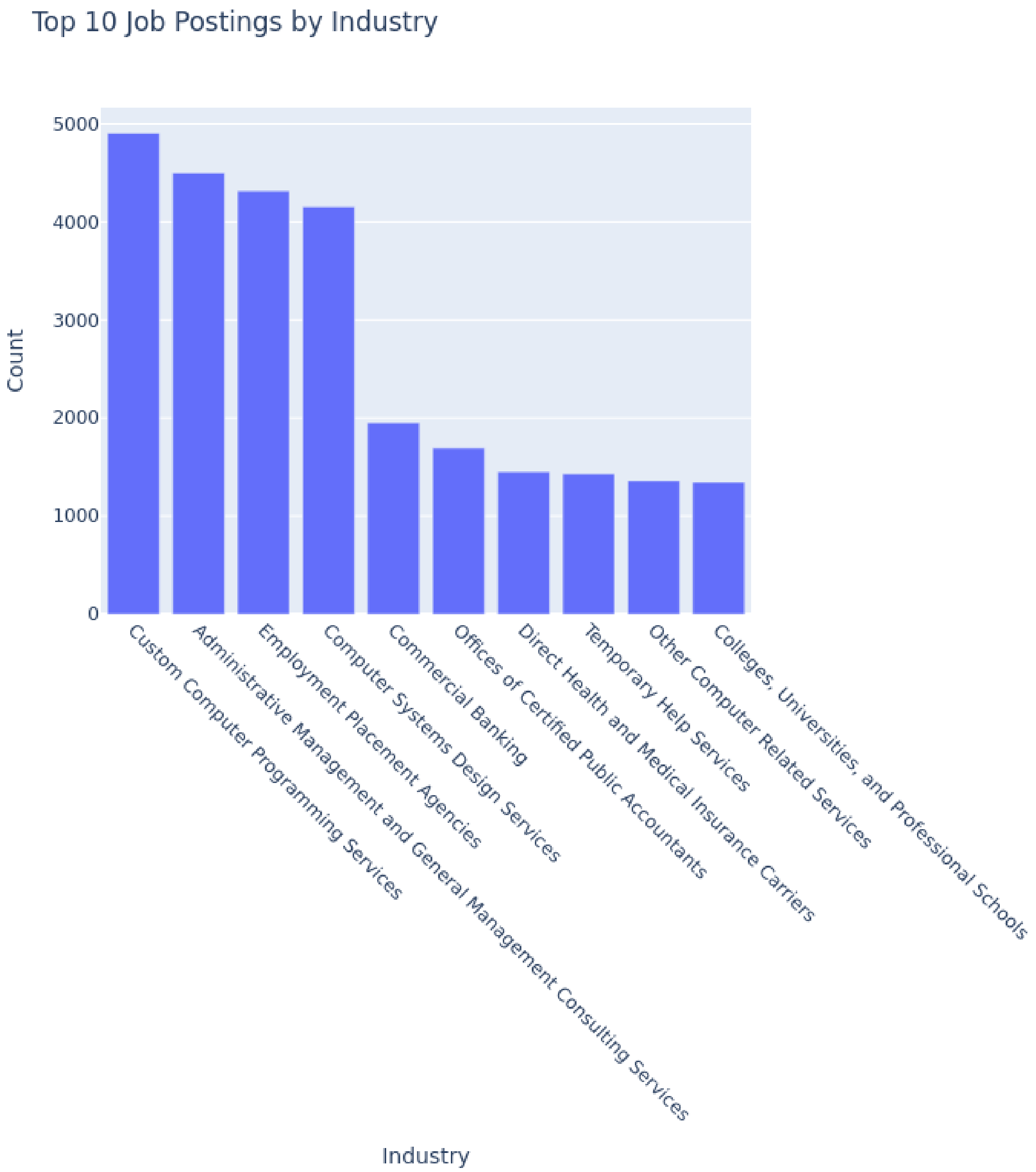


The pie chart represents the distribution of remote, on-site, and hybrid job postings. It helps visualize the proportion of different work arrangements in the job market.

Exploratory Data Analysis

Enhanced Visuals

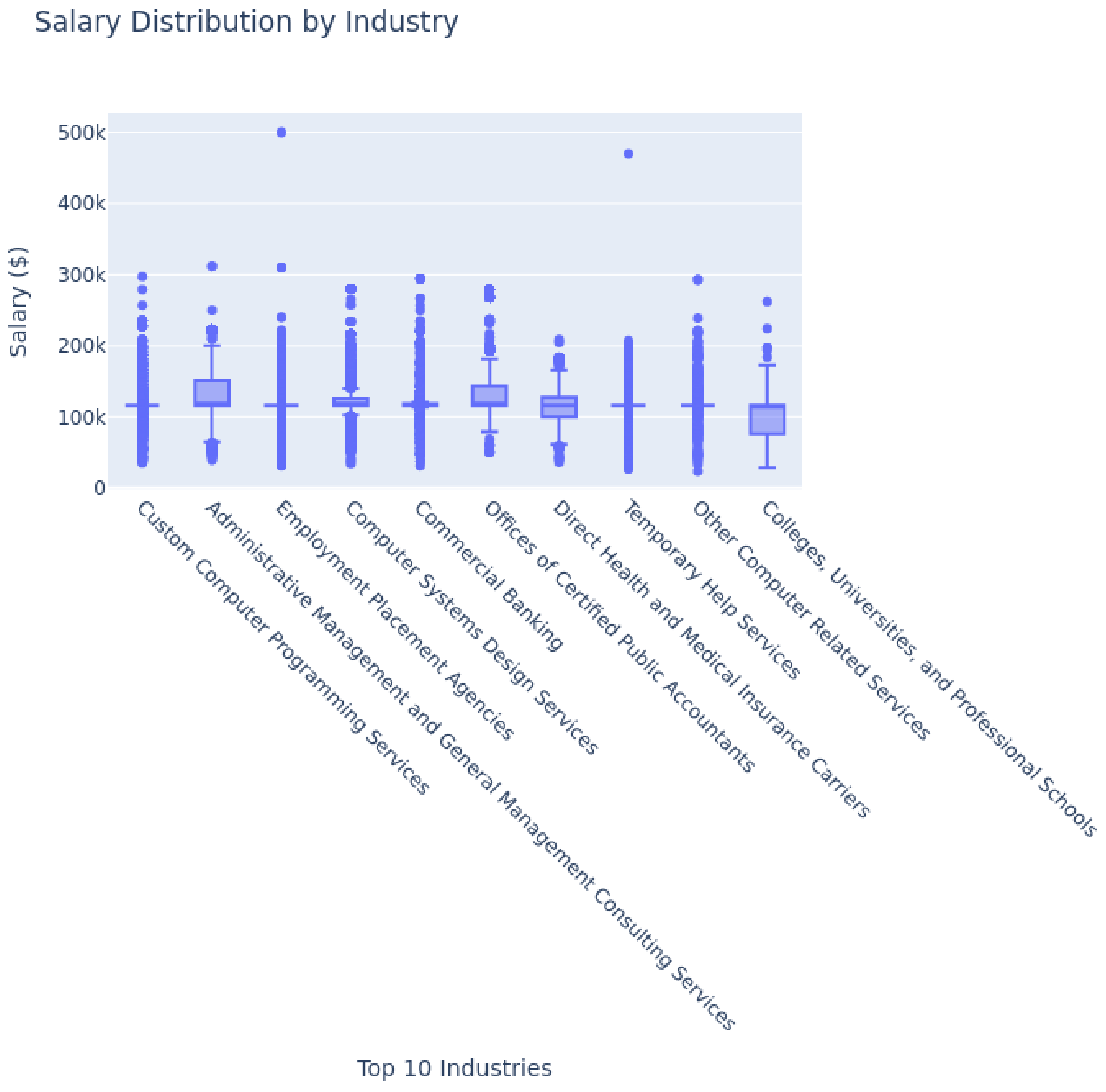
Top 10 Job Postings by Industry



> The most frequently advertised job postings come from Custom Computer Programming Services, Accounting Services, and Employment Placement Agencies. These industries are

consistently hiring across roles, suggesting a high demand for software developers, finance professionals, and recruiters. This indicates strong hiring momentum in tech and support functions.

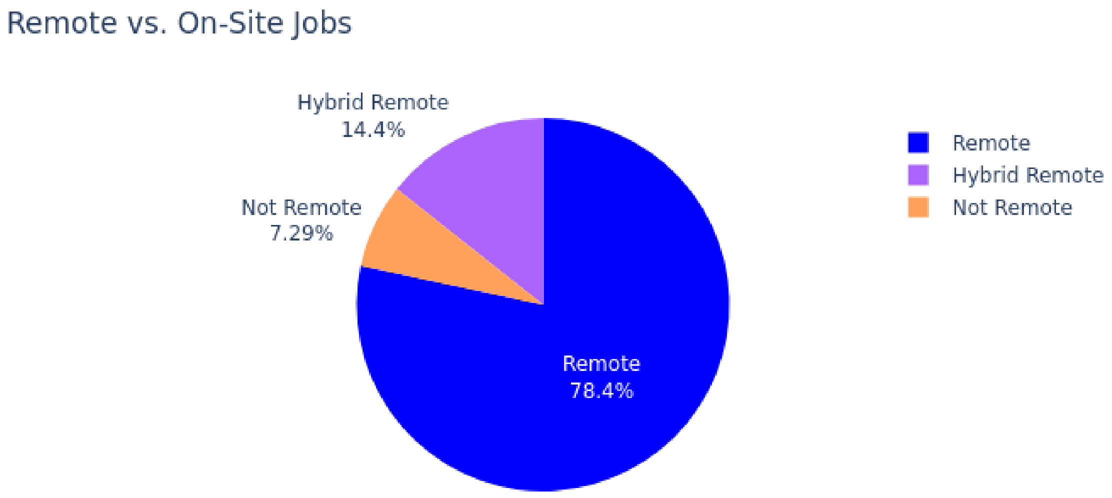
Salary Distribution by Industry



> Salary distribution varies widely across industries. While most sectors show a median salary between $80K and $120K, certain fields like Commercial Banking and Offices of Certified

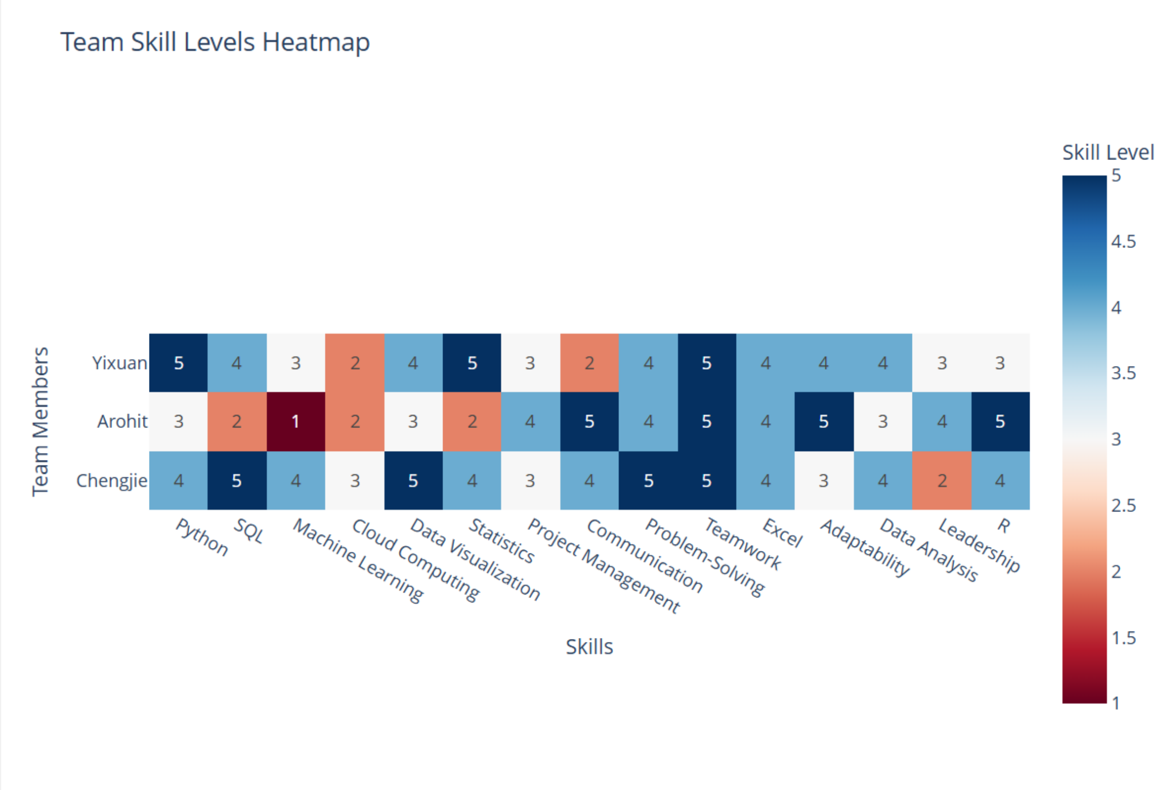
Public Accountants show higher outliers, indicating potential for high-earning roles. The variation within each industry also reflects differing job levels and skill demands.

Remote vs. On-Site Jobs



> Over 78% of job postings offer remote work options, either fully or in hybrid mode. This highlights the growing normalization of flexible work arrangements post-pandemic. Only 7% of jobs are strictly on-site, indicating a permanent shift in job design and workplace expectations.

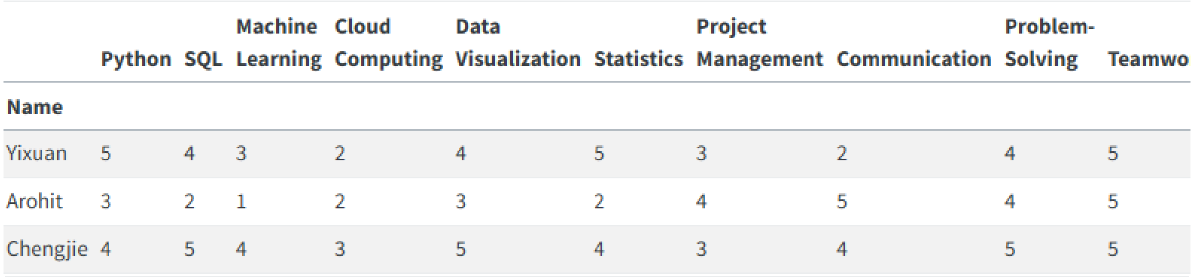
Team Skill Levels Heatmap



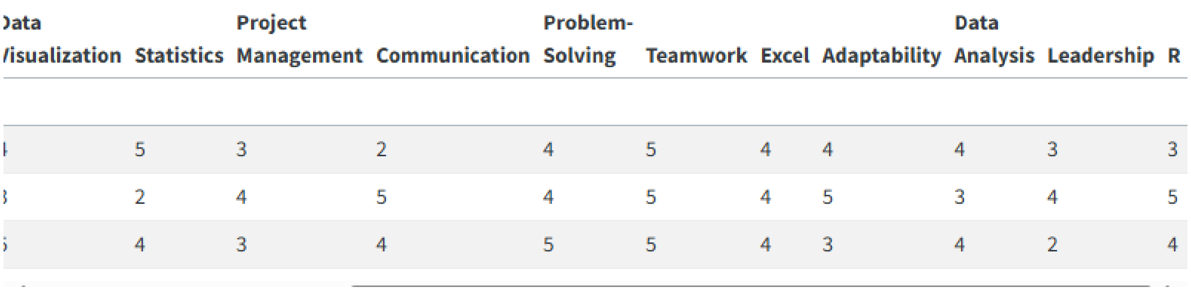
> The team demonstrates strong skill levels in Communication, Problem-Solving, and Teamwork, all scoring 5 across members. However, there are visible gaps in Machine Learning and Cloud Computing, particularly for Arohit. These gaps highlight potential areas for upskilling to align with industry demands in data and engineering roles.

Skill Gap Analysis

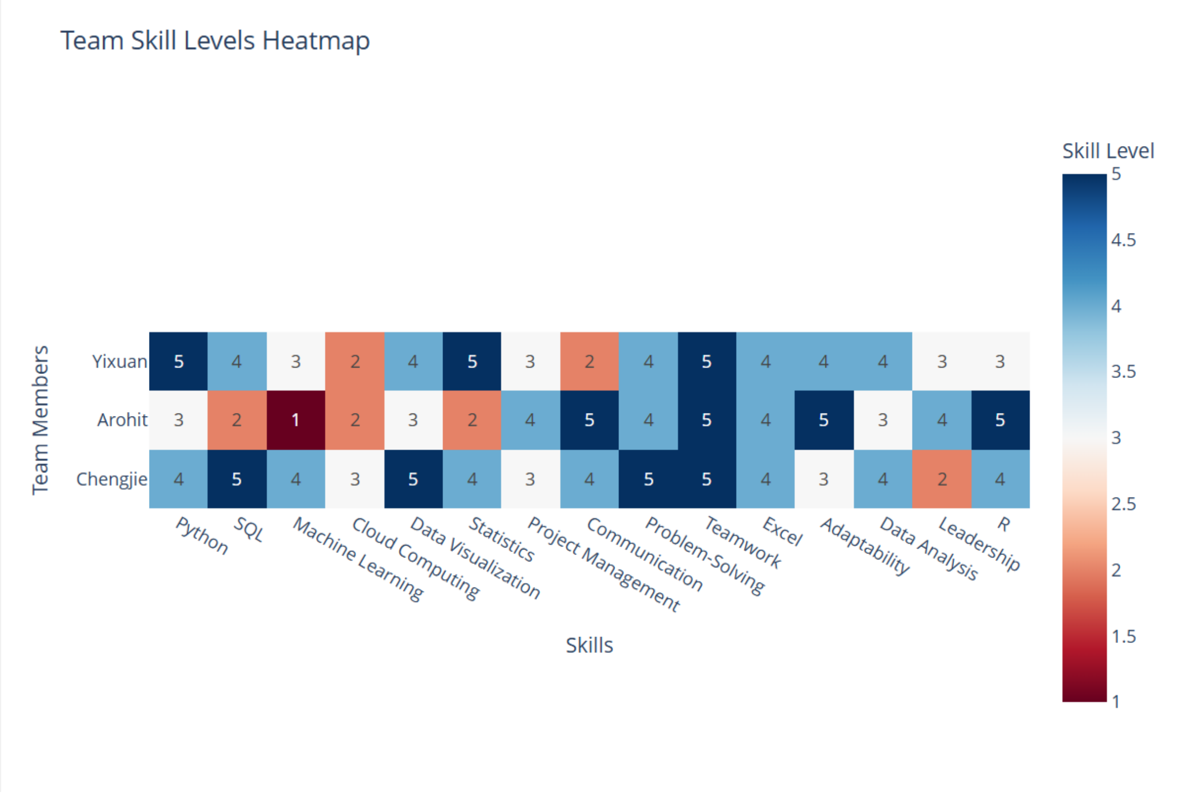
	Python	SQL	Machine Learning	Cloud Computing	Data Visualization	Statistics	Project M
Name							
Yixuan	5	4	3	2	4	5	3
Arohit	3	2	1	2	3	2	4
Chengjie	4	5	4	3	5	4	3



	Python	SQL	Machine Learning	Cloud Computing	Data Visualization	Statistics	Project Management	Communication	Problem-Solving	Teamwo
Name										
Yixuan	5	4	3	2	4	5	3	2	4	5
Arohit	3	2	1	2	3	2	4	5	4	5
Chengjie	4	5	4	3	5	4	3	4	5	5



Data Visualization	Statistics	Project Management	Communication	Problem-Solving	Teamwork	Excel	Adaptability	Data Analysis	Leadership	R
4	5	3	2	4	5	4	4	4	3	3
3	2	4	5	4	5	4	5	3	4	5
5	4	3	4	5	5	4	3	4	2	4



Personalized Learning Plan

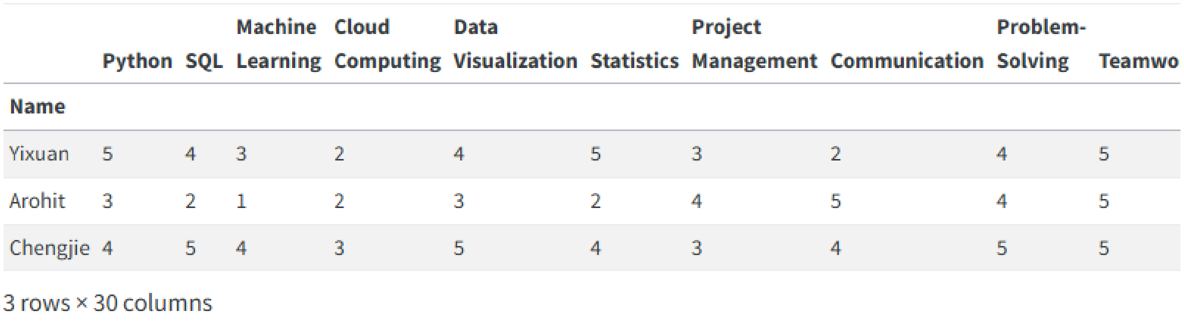
Based on the heatmap and extracted job skill requirements, the following areas are recommended for improvement:

- **Yixuan**: Should focus on improving **Communication** and **Cloud Computing**, which are below average and frequently required by employers.
- **Arohit**: Needs significant upskilling in **Machine Learning**, **Statistics**, and **Data Visualization**, which are critical for data-centric roles.
- **Chengjie**: Should enhance **Leadership** and **Adaptability** skills, which are essential for project coordination and dynamic environments.

Courses on platforms such as **Coursera**, **edX**, or **LinkedIn Learning** can be recommended to address these gaps effectively.

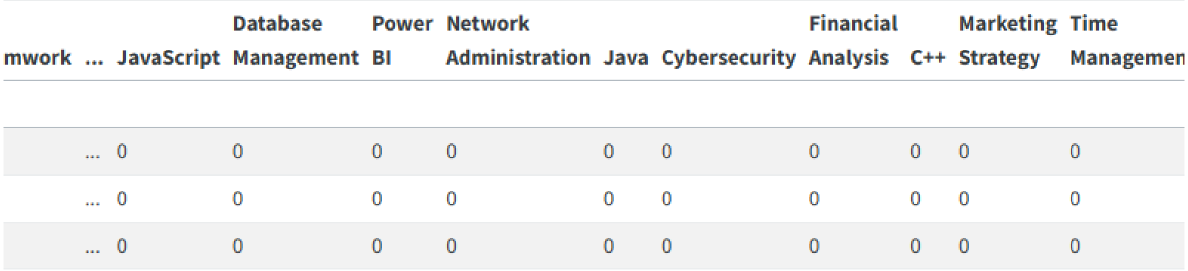
```
{'Data Visualization', 'C++', 'Time Management', 'HTML/CSS', 'Leadership', 'Java', 'Teamwork
```

	Python	SQL	Machine Learning	Cloud Computing	Data Visualization	Statistics	Project M
Name							
Yixuan	5	4	3	2	4	5	3
Arohit	3	2	1	2	3	2	4
Chengjie	4	5	4	3	5	4	3



	Python	SQL	Machine Learning	Cloud Computing	Data Visualization	Statistics	Project Management	Communication	Problem-Solving	Teamwo
Name										
Yixuan	5	4	3	2	4	5	3	2	4	5
Arohit	3	2	1	2	3	2	4	5	4	5
Chengjie	4	5	4	3	5	4	3	4	5	5

3 rows × 30 columns



mwork	...	JavaScript	Database Management	Power BI	Network Administration	Java	Cybersecurity	Financial Analysis	C++	Marketing Strategy	Time Managemen
	...	0	0	0	0	0	0	0	0	0	0
	...	0	0	0	0	0	0	0	0	0	0
	...	0	0	0	0	0	0	0	0	0	0

Conclusion

This skill gap analysis reveals critical strengths in collaboration, problem-solving, and communication within the team. However, technical gaps—particularly in Machine Learning, Cloud Computing, and Leadership—need to be addressed to align with job market demands. The personalized learning plans are tailored to ensure all members enhance relevant skills for competitive employability in the data and tech sectors.

1. Personalized Learning Plan

Based on our skill gap analysis, we developed individual development goals to enhance each team member's capabilities. The heatmap and table showed skill levels across 15 competencies including Python, Machine Learning, Communication, and Leadership.

Yixuan Yang

- **Strengths**: Python (5), Problem-Solving (4), Teamwork (5), Leadership (3)
- **Development Focus**:
  - Cloud Computing (2): Take online labs (e.g., AWS Academy, GCP Fundamentals)
  - Communication (2): Practice presentations and join group discussions weekly
  - Project Management (3): Consider introductory PM courses (Coursera or LinkedIn Learning)

Arohit Talari

- **Strengths**: Communication (5), Teamwork (5), Leadership (4), R (5)
- **Development Focus**:
  - SQL (1): Complete SQL bootcamp on DataCamp
  - Machine Learning (2): Study basics with scikit-learn tutorials
  - Data Visualization (2): Learn with Tableau or Plotly tutorials
  - Cloud Computing (2): Follow step-by-step GCP training

Chengjie Lu

- **Strengths**: Python (4), Data Visualization (5), SQL (5), Communication (5)
- **Development Focus**:
  - Leadership (2): Take initiative in meetings, and read leadership case studies
  - Adaptability (2): Volunteer for cross-functional tasks or new tool trials

2. Actionable Timeline

Member	Focus Area(s)	Resource	Target Date
Yixuan	Cloud Computing, PM	AWS Labs, Coursera Intro PM	June 2025
Arohit	SQL, ML, Viz	DataCamp, scikit-learn, Tableau	June 2025
Chengjie	Leadership, Adaptability	Case studies, team role rotation	June 2025

By tailoring development plans to individual gaps, we aim to strengthen our team's readiness for data-driven projects and leadership roles in 2024 and beyond.

Multiple Linear Regression - Salary Predition

Mutiple Linear Regression

	LAST_UPDATED_DATE	POSTED	EXPIRED	DURATION	SOURCE_TYPES	SOURCES
0	2024-09-06	2024-06-02	2024-06-08	6.0	[\n ”Company”\n]	[\n ”brass
1	2024-08-02	2024-06-02	2024-08-01	-1.0	[\n ”Job Board”\n]	[\n ”maine
2	2024-09-06	2024-06-02	2024-07-07	35.0	[\n ”Job Board”\n]	[\n ”dejob
3	2024-09-06	2024-06-02	2024-07-20	48.0	[\n ”Job Board”\n]	[\n ”disab
4	2024-06-19	2024-06-02	2024-06-17	15.0	[\n ”FreeJobBoard”\n]	[\n ”craigs

Feature Engineering

```
(69199, 60)
exp_mid                                         float64
MODELED_DURATION                                float64
skill_count                                       int64
has_python                                        int64
edu_ge_bachelors                                  int64
SALARY                                          float64
EMPLOYMENT_TYPE_NAME_Part-time (  32 hours)    float64
EMPLOYMENT_TYPE_NAME_Part-time / full-time      float64
REMOTE_TYPE_NAME_Not Remote                     float64
REMOTE_TYPE_NAME_Remote                         float64
dtype: object
```

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

```
count     20760.000000
mean     116858.597141
std        8723.766929
min       88939.521987
25%      110186.143187
50%      115748.686133
75%      122510.993586
max      163152.472089
dtype: float64
```

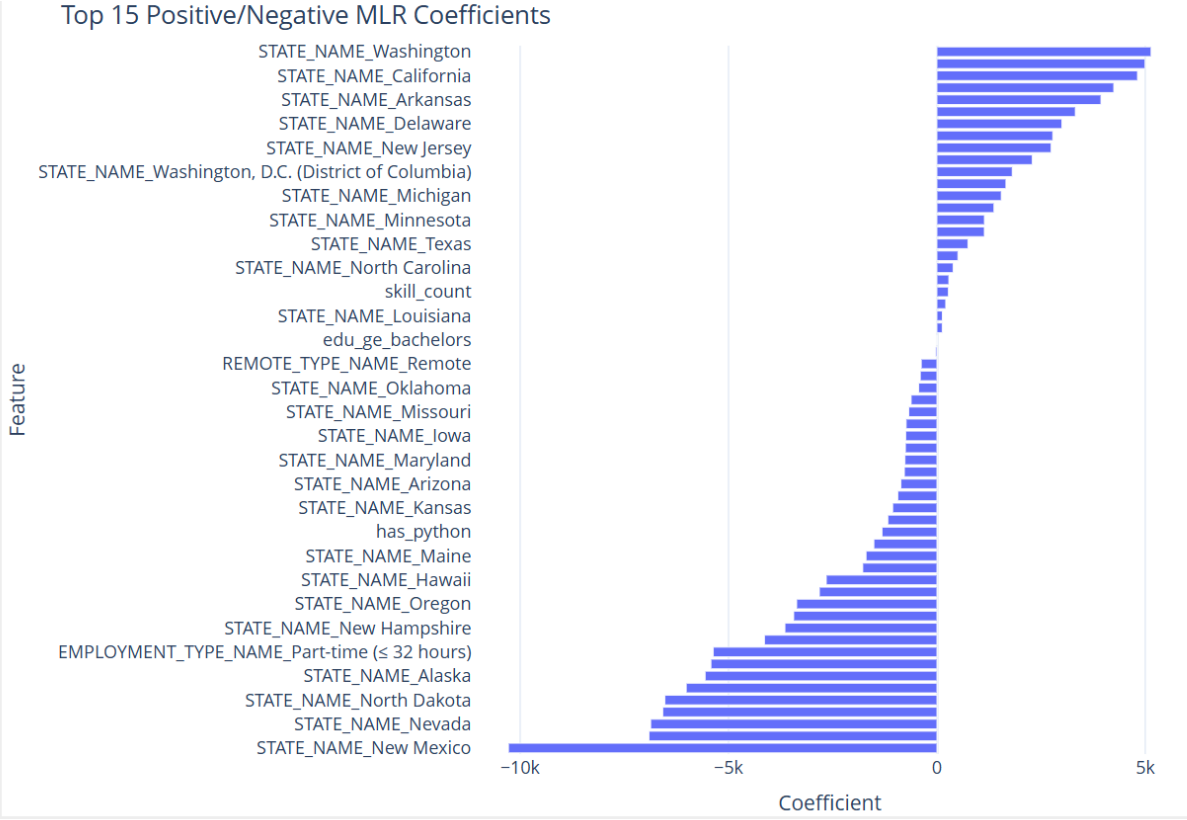
```
RMSE: 777720416.43
R-squared: 0.0874
```

	Feature	Coefficient
54	STATE_NAME_Washington	5135.856412
52	STATE_NAME_Vermont	4992.125928
12	STATE_NAME_California	4810.124902
14	STATE_NAME_Connecticut	4240.562772
11	STATE_NAME_Arkansas	3933.241423
0	exp_mid	3319.050797
15	STATE_NAME_Delaware	2997.204286
20	STATE_NAME_Illinois	2777.969965
37	STATE_NAME_New Jersey	2739.092860
53	STATE_NAME_Virginia	2287.668518

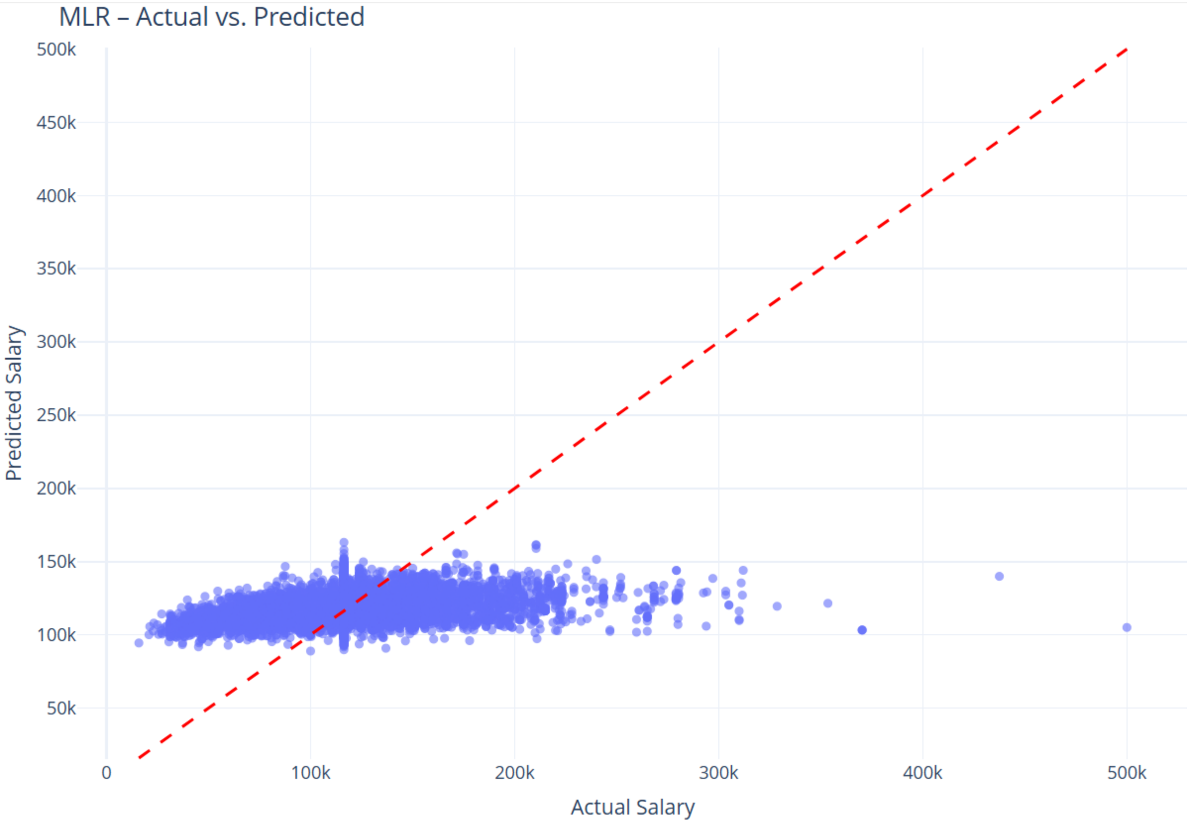
	Feature	Coefficient
54	STATE_NAME_Washington	5135.856412
52	STATE_NAME_Vermont	4992.125928
12	STATE_NAME_California	4810.124902
14	STATE_NAME_Connecticut	4240.562772
11	STATE_NAME_Arkansas	3933.241423
0	exp_mid	3319.050797
15	STATE_NAME_Delaware	2997.204286
20	STATE_NAME_Illinois	2777.969965
37	STATE_NAME_New Jersey	2739.092860
53	STATE_NAME_Virginia	2287.668518

Visualization

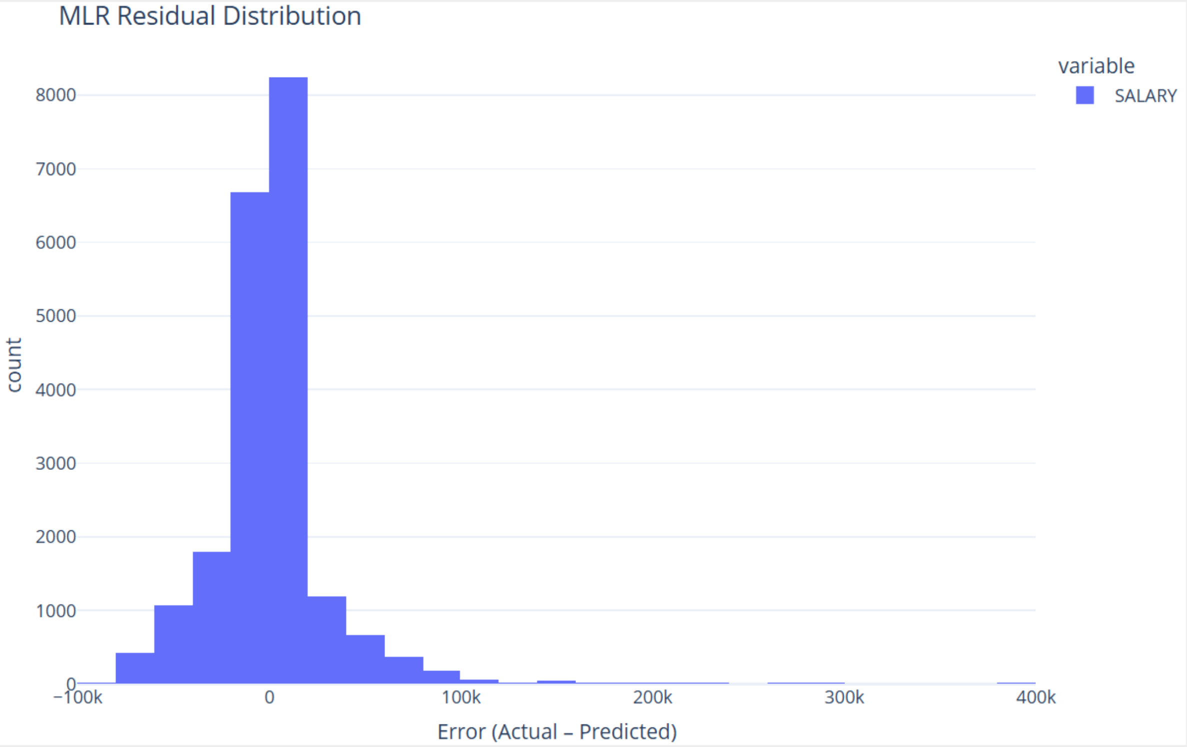
Coefficient bar chart



Actual vs. Predicted



Residual histogram

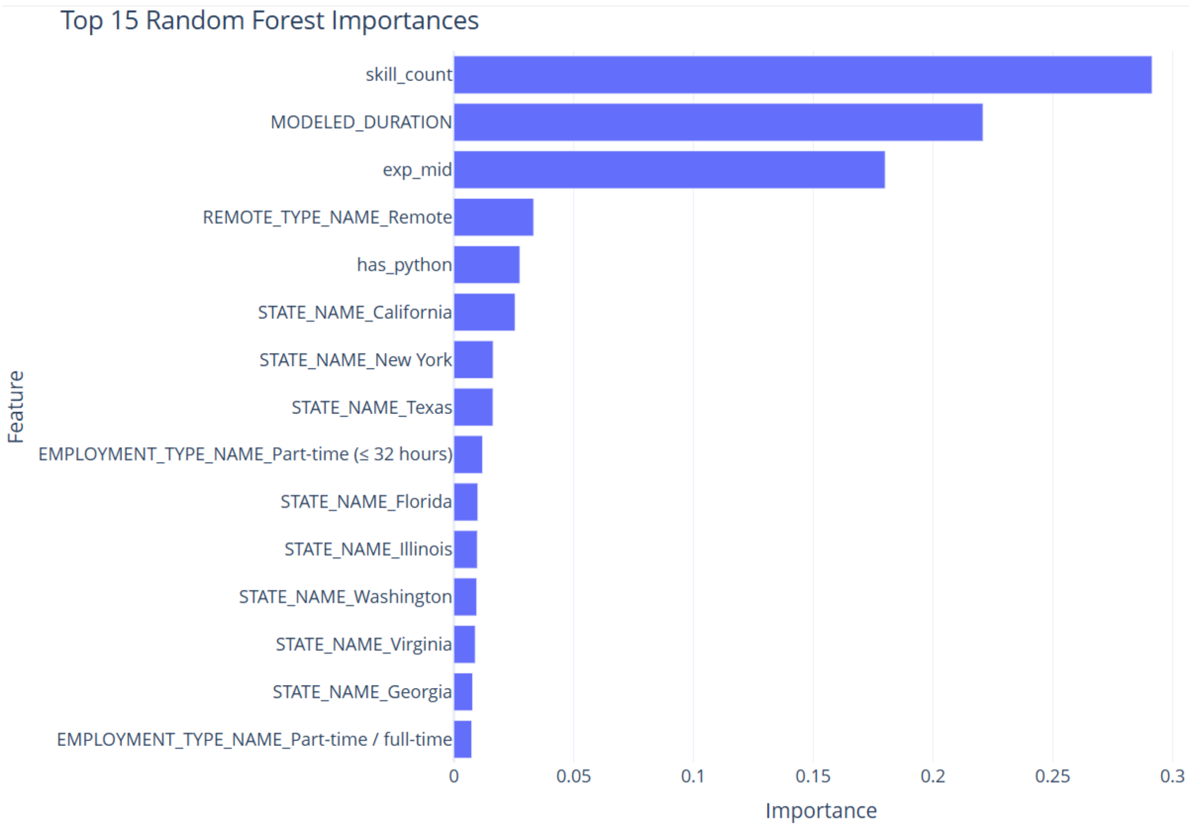


Random Forest

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

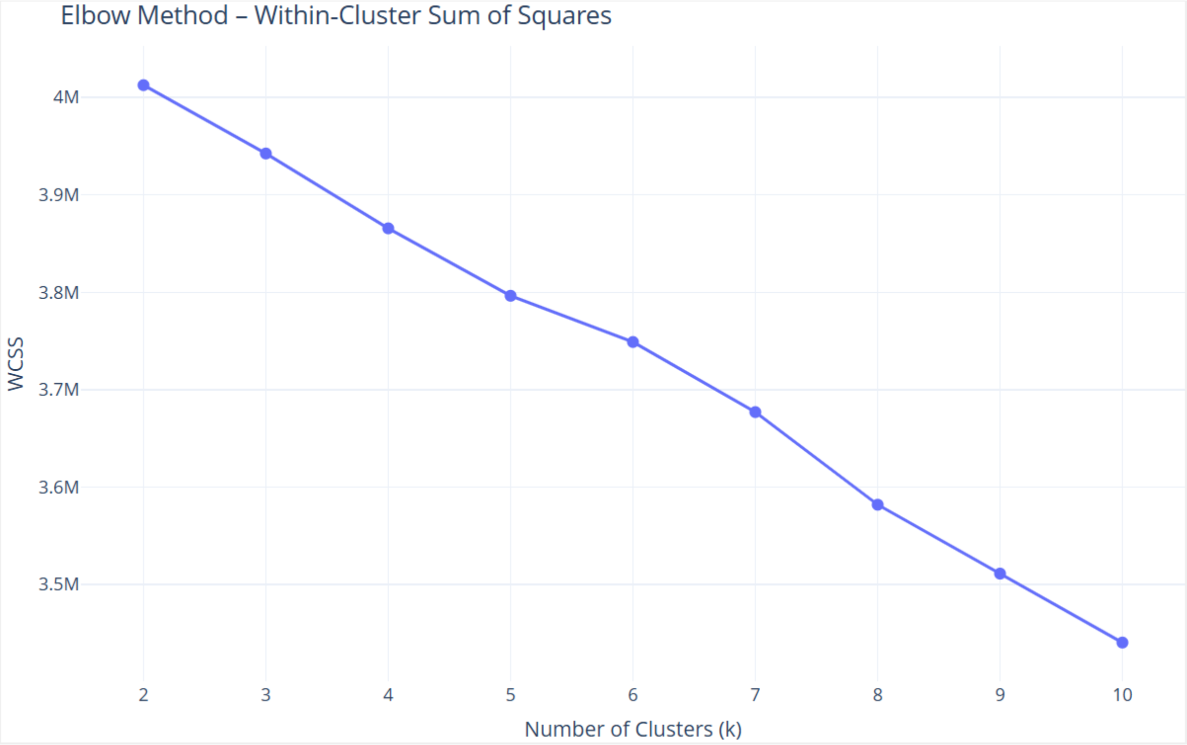
```
Random Forest  •  RMSE = 637552503.24  |  R² = 0.252
```

Rank Importance

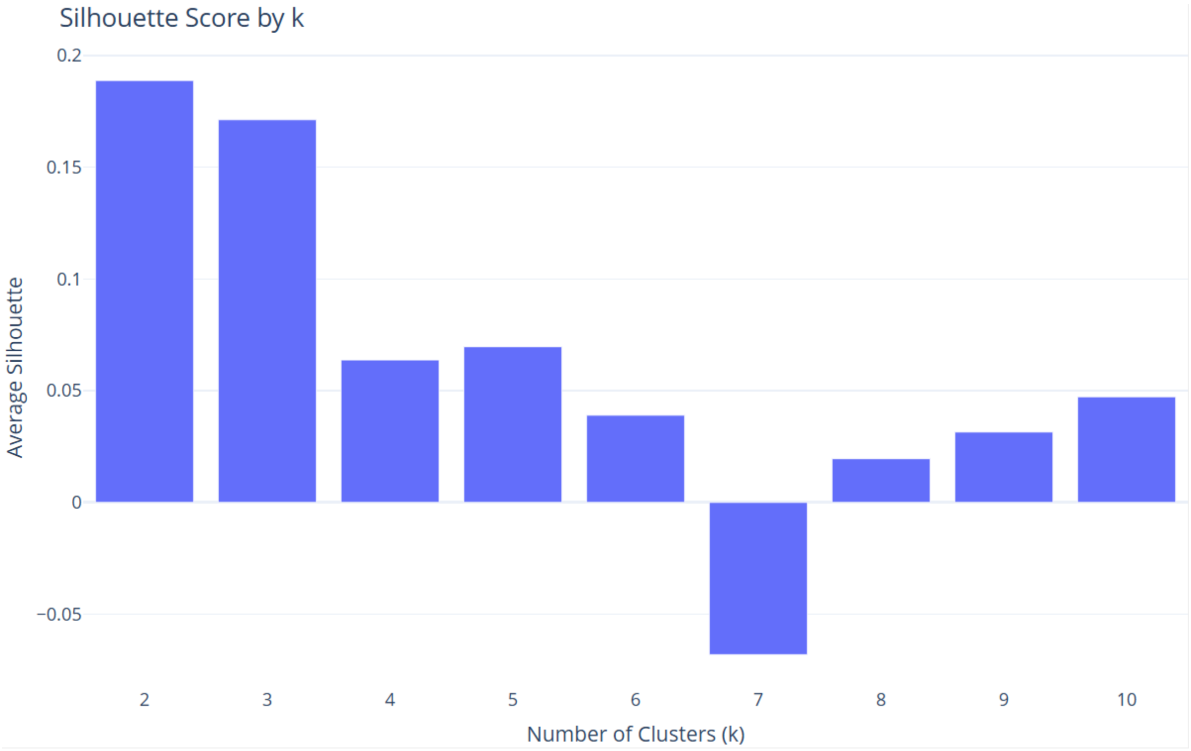


Unsupervised Learning - Kmeans

Elbow Plot



Silhouette Score



Multiple Linear Regression

In this section, we built a multiple linear regression model to predict salaries using a variety of features, including experience, skill counts, education, and employment type.

Key results: - **RMSE**: 77724.06 - **R-squared**: 0.0874

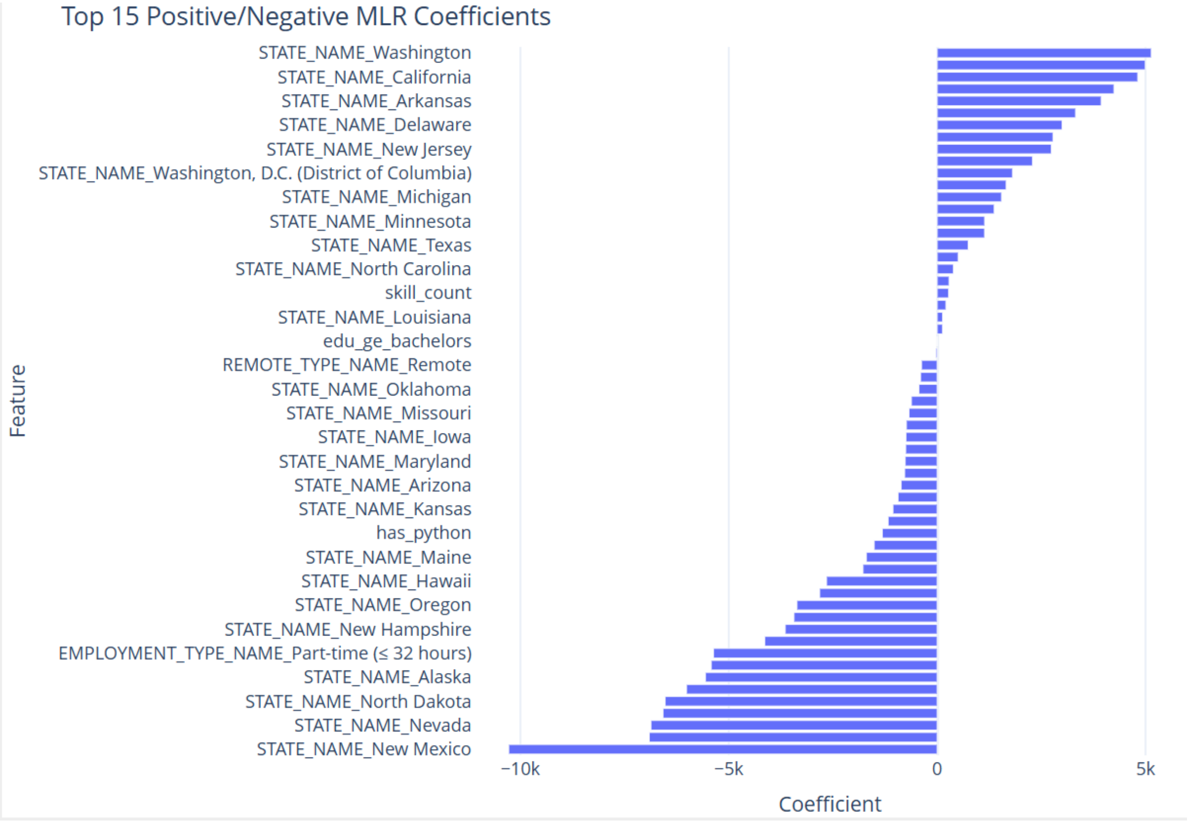
These metrics indicate that while the model provides some insight, there is significant unexplained variance, suggesting that salary prediction is complex and influenced by additional unobserved factors.

Top features influencing salary (positive coefficients): - STATE_NAME_Washington (+$5135.86) - STATE_NAME_Vermont (+$4992.13) - STATE_NAME_California (+$4810.12) - STATE_NAME_Connecticut (+$4240.56)

The results show that the location (state) plays a crucial role in determining salary.

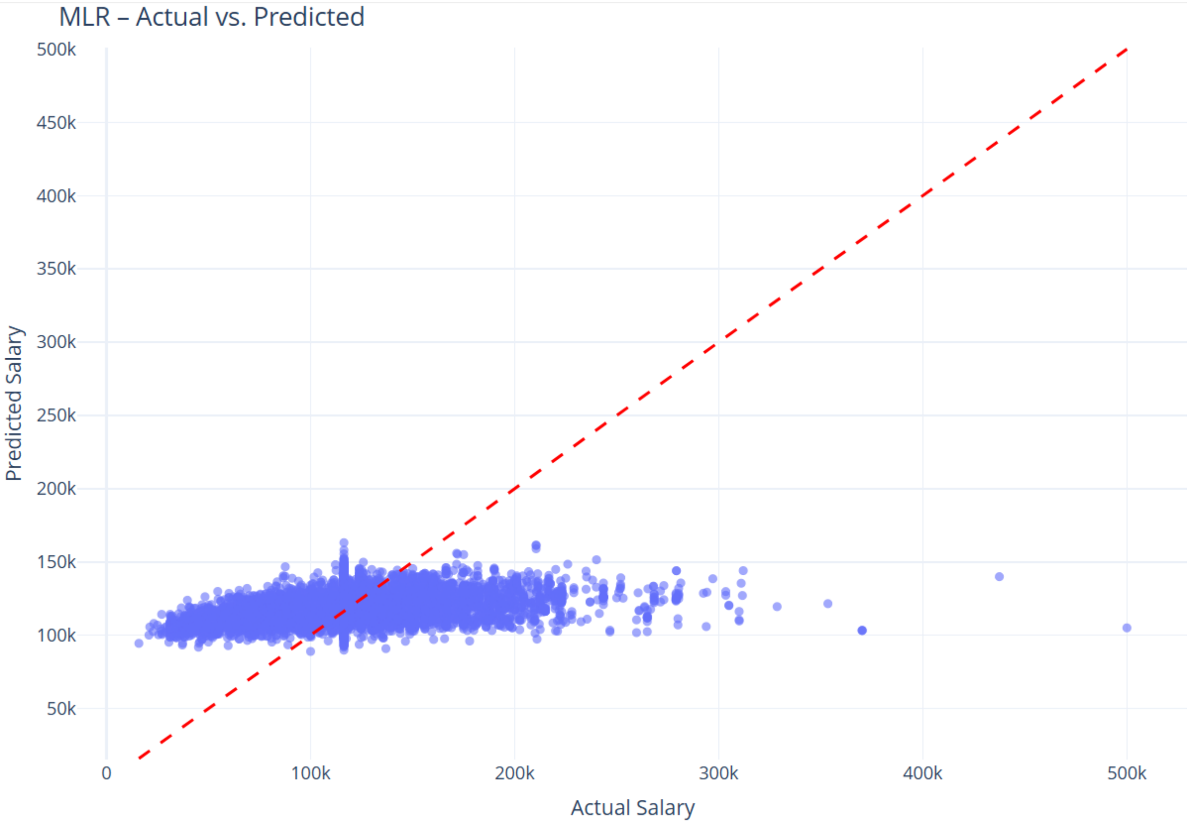
Visualizations

Coefficient Bar Chart



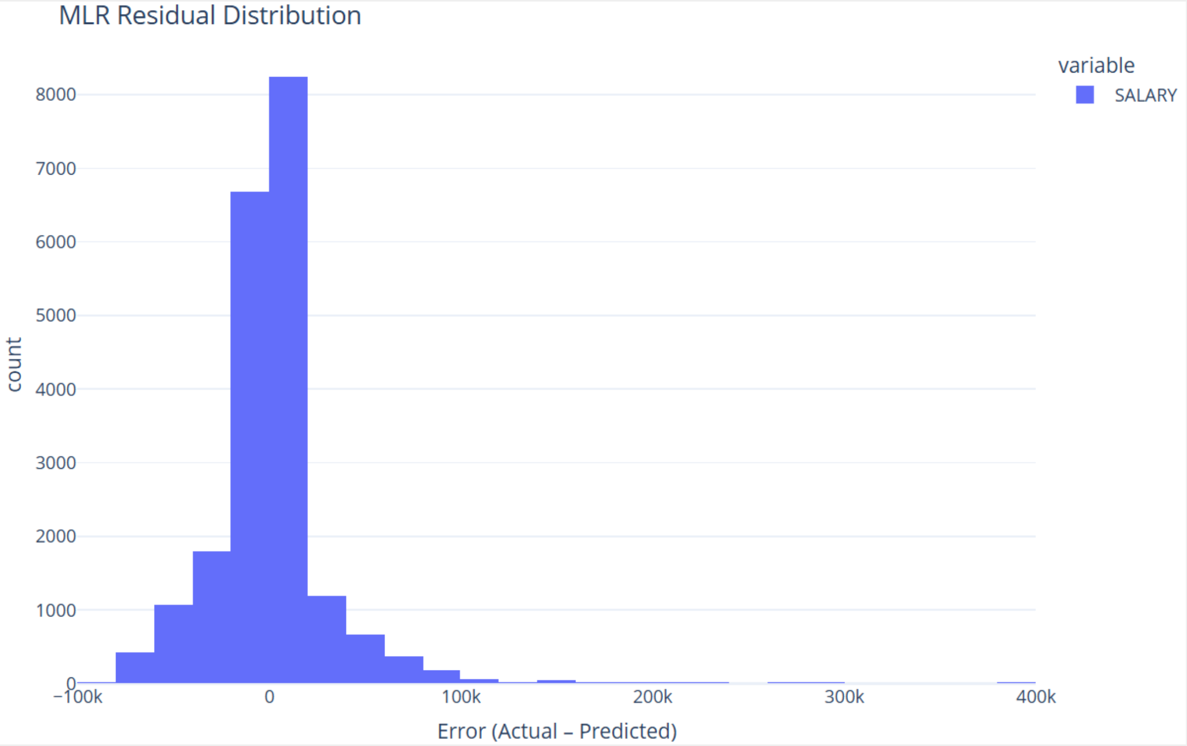
A bar chart was used to visualize the top positive and negative influences on predicted salaries. Positive coefficients primarily relate to states with higher living costs.

Actual vs. Predicted Plot



A scatter plot comparing actual salaries against predicted salaries shows a wide dispersion, indicating prediction inaccuracies at extreme salary values.

Residual Histogram



The histogram of residuals suggests a concentration of errors near zero but with some large deviations, reinforcing the need for model improvement.

Random Forest Regression

A random forest model was implemented to enhance prediction performance.

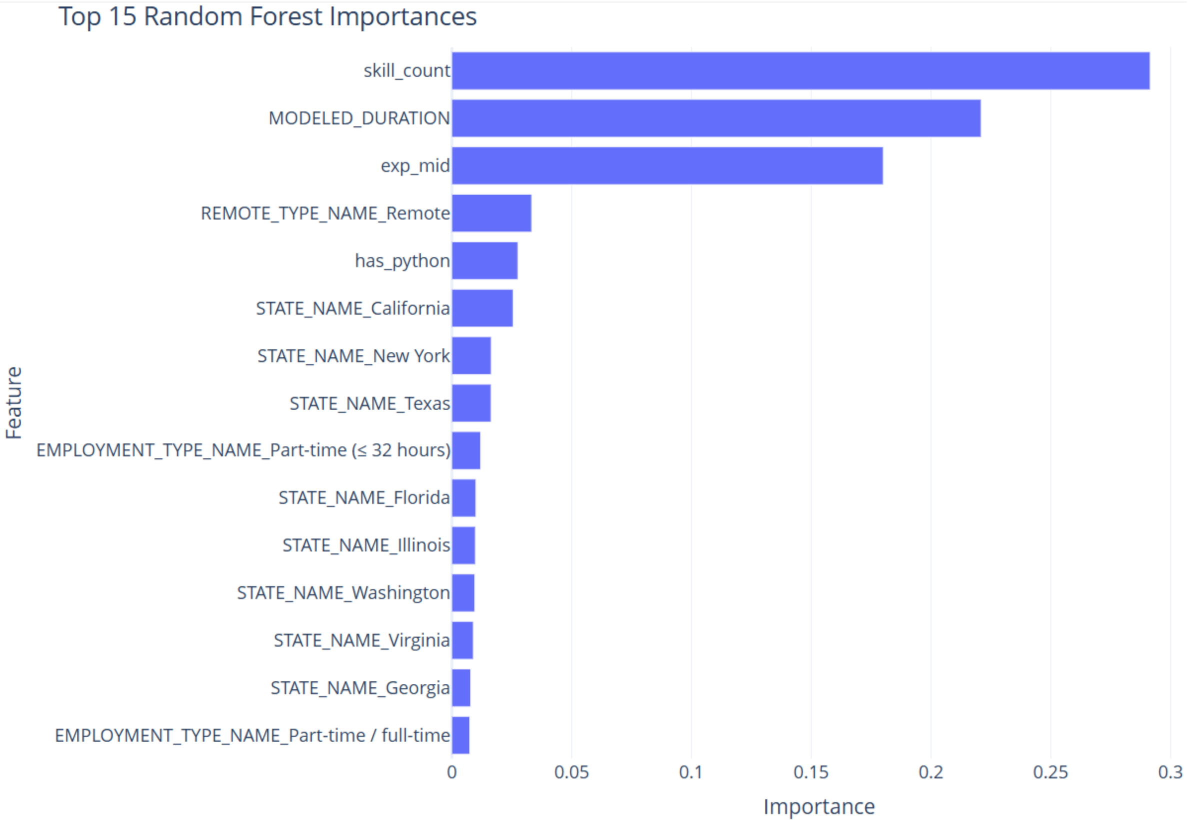
Key results: - **RMSE**: 63755.28 - **R-squared**: 0.252

Compared to multiple regression, random forest achieves better fit, although a large proportion of variance still remains unexplained.

Top features by importance: - skill_count - MODELED_DURATION - exp_mid - REMOTE_TYPE_NAME_Remote

Skill counts and modeled duration (likely representing experience) have the highest impact on salary predictions.

Rank Importance Chart

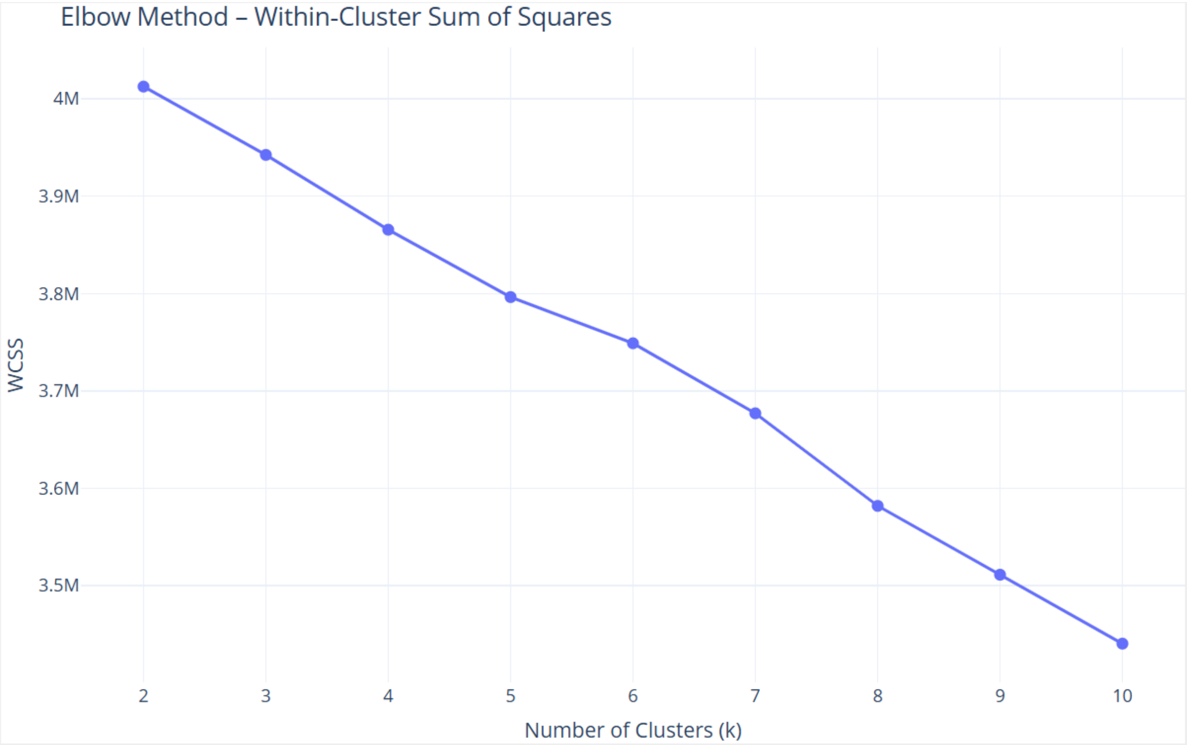


A bar chart displays the relative importance of the top 15 features, confirming the key role of skills and experience.

Unsupervised Learning: KMeans Clustering

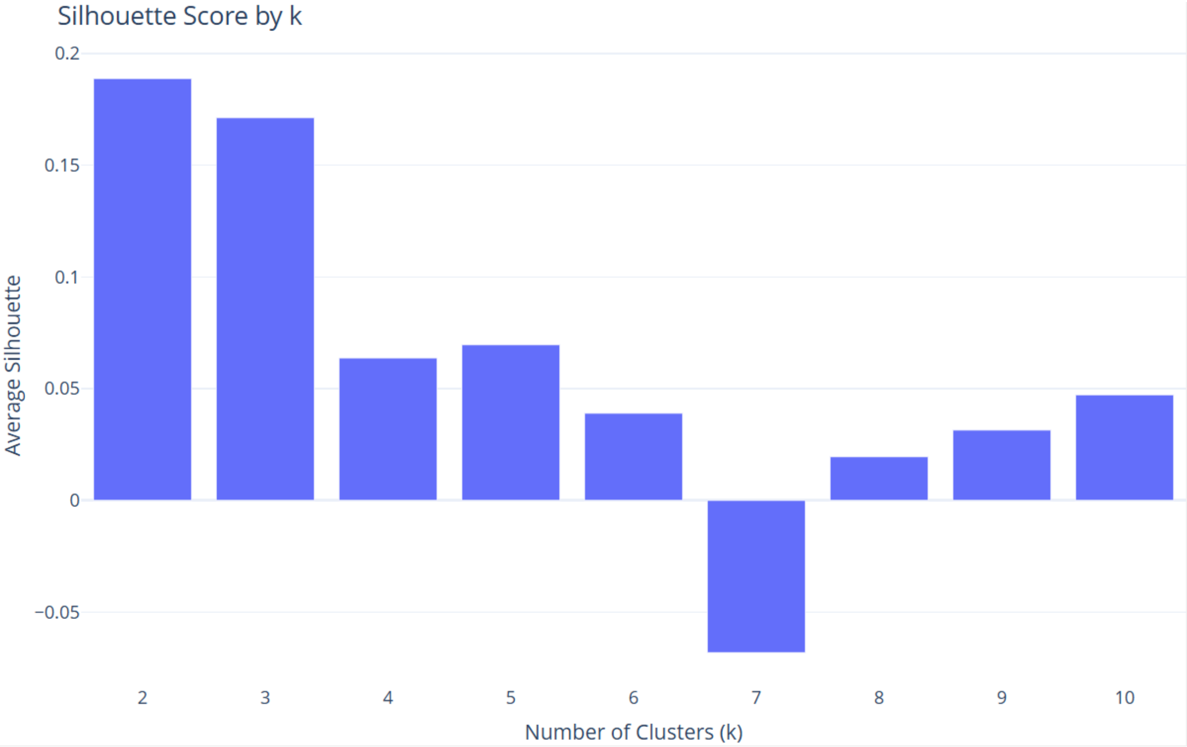
KMeans clustering was used to segment jobs into different clusters based on attributes.

Elbow Plot



The elbow plot suggests that an optimal number of clusters is likely around **3** to **4**, as the rate of decrease in within-cluster sum of squares slows beyond this point.

Silhouette Score



Silhouette scores were plotted for different numbers of clusters. The highest silhouette score is achieved at **k=2** (approximately 0.19), suggesting that two clusters provide the clearest separation.

Conclusion

This analysis demonstrates that predicting salary is highly complex. Multiple regression and random forest models reveal that factors such as experience, skill count, and location significantly impact salary. Clustering analysis shows that job characteristics can be segmented into relatively clear groups, though the differences between some clusters are subtle.

Future work should include exploring additional variables (e.g., company size, industry sector) and more advanced modeling techniques (e.g., gradient boosting, neural networks) to improve prediction accuracy.

Natural Language Processing (NLP) Analysis

Introduction

In this section, we conduct a basic Natural Language Processing (NLP) analysis based on job descriptions in our dataset (`cleaned_job_postings.csv`). The goal is to extract key topics and skills mentioned in job postings, enhancing our understanding of employer expectations in the market.

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

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

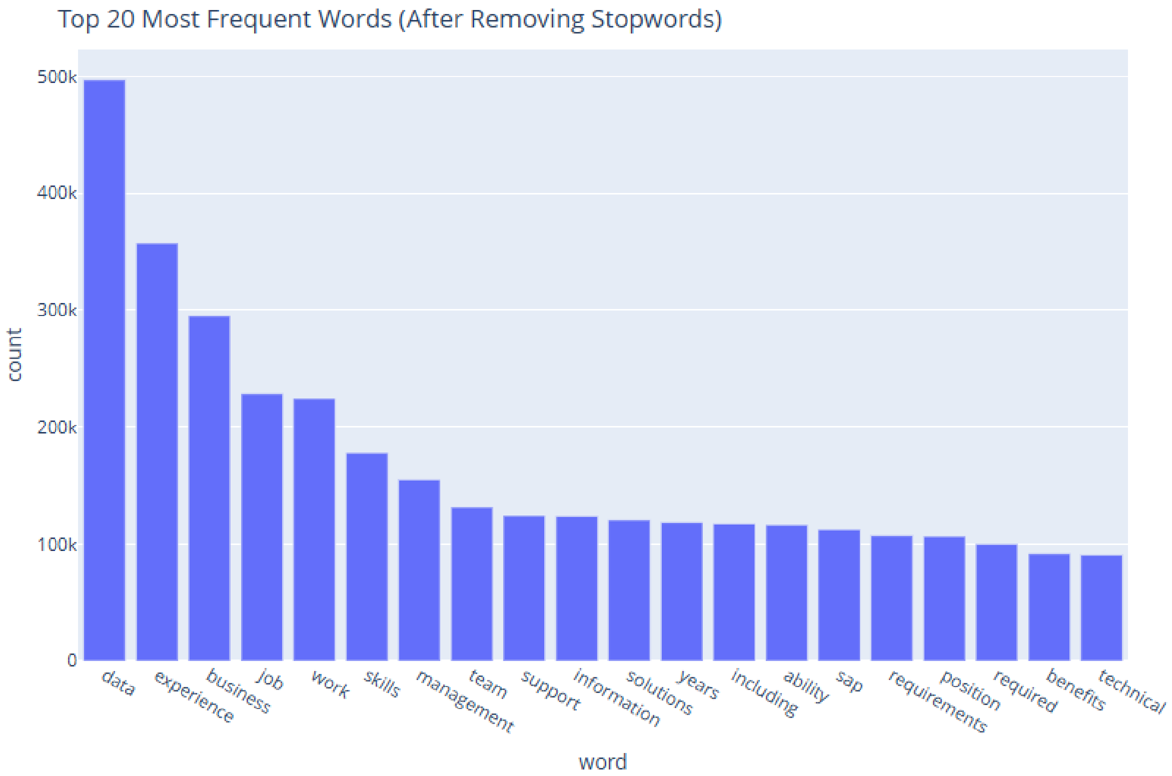
```
Topic 1:
work | health | data | information | benefits | required | position | insurance | time | empl

Topic 2:
business | work | team | benefits | technology | skills | status | solutions | data | compan

Topic 3:
business | clients | oracle | work | sap | range | employment | role | solutions | client

Topic 4:
sap | business | management | oracle | skills | requirements | solutions | technical | years

Topic 5:
data | business | skills | analysis | work | analyst | analytics | management | ability | to
```



Topic Modeling Summary

Based on the LDA topic modeling results, we can identify five major underlying themes within the job descriptions:

- **Topic 1**: Focuses on **work, health, data, benefits, insurance, and employment terms**, suggesting that many job postings emphasize hiring conditions, employee benefits, and information handling requirements.
- **Topic 2**: Centers around **teamwork, technology, and skill development**, highlighting the growing importance of collaboration and technical proficiency in hiring practices.
- **Topic 3**: Emphasizes **client management, Oracle systems, and employment roles**, reflecting strong demand for CRM (Customer Relationship Management) and ERP (Enterprise Resource Planning) system skills.
- **Topic 4**: Highlights **SAP expertise, management skills, and technical requirements**, indicating a sustained need for advanced management and systems integration capabilities.

- **Topic 5**: Concentrates on **data analytics, business intelligence, and skill applications**, showing a strong market preference for data-driven decision-making and analytical roles.

Overall Interpretation

Current job postings repeatedly emphasize **technical competencies** (such as **Oracle** and **SAP**), **data analytics capabilities**, and **cross-functional communication skills**. Additionally, **benefits**, **health insurance**, and specific **employment requirements** are critical elements emphasized by employers during recruitment.

1. Summary

This project integrates market trend analysis, skill benchmarking, and machine learning modeling to assess personal job readiness. Through rigorous data exploration and predictive modeling, we uncovered several important insights:

- **Industry Trends**: High demand is concentrated in the technology, consulting, and support service industries.
- **Salary Drivers**: Salary disparities are largely influenced by years of experience, the number of skills possessed, and geographic location.
- **Team Skill Assessment**: Team strengths are notable in areas of communication and problem-solving. However, skill gaps were identified in **cloud computing** and **machine learning** competencies.

These findings provide a grounded view of current labor market expectations and highlight actionable areas for professional development.

2. Future Directions

Looking ahead, several strategic pathways emerge:

- **Skill Development**: By implementing the personalized learning plans and focusing on closing the identified skill gaps, each member can substantially enhance their employability.
- **Continuous Trend Monitoring**: Staying attuned to evolving industry demands will ensure alignment between skills and market needs over time.
- **Scalability of Methods**: The analytic framework and techniques developed in this project are adaptable. They can be scaled to assist broader populations of job seekers and can be applied to a variety of career planning and workforce development scenarios.

Overall, this project demonstrates a robust approach to data-driven career readiness evaluation, setting a strong foundation for future strategic personal development.