Job Market Research 2024

Final Report

Angelina McKim

Devin Blanchard

Leo Liu

2025-10-10

Table of contents

# 1. Introduction & Research Rationale

The job market in 2024 is undergoing a major transformation with the rise of artificial intelligence and widespread acceptance of long-distance/remote work. This analysis will examine the following key topics to uncover the factors that influence salary trends:

**The restructuring of compensation due to AI, inflation, and remote work.**

The traditional pay structure is being reshaped as companies adapt to automation, rising costs of living, and a distributed workforce. These forces are pushing employers to rethink how they reward skills, experience, and location.

**Growing pay disparities across regions, industries, and job types in 2024.**

High-paying jobs are increasingly concentrated in specific tech-driven sectors and urban hubs. Meanwhile, many essential or service-based roles are seeing slower wage growth, deepening inequality across the workforce.

**Shifting salary patterns based on remote flexibility, job type, and sector growth.**

Remote roles now often offer comparable or even higher compensation due to talent shortages and broader applicant pools. Industries like tech and healthcare are setting new standards, while others struggle to keep pace.

# 2. Literature Review Summary

The rise of Artificial Intelligence (AI) has reshaped the job market across the US, with demand for AI-related skills increasing dramatically. According to a study by PWC, jobs that require AI specialist skills are growing 3.5 times faster than all other job markets, with skilled AI workers being paid up to 25% more in some sectors (PwC 2024). However, this increase in compensation is not limited to workers specialized in AI. Non-AI roles that require complementary skills such as digital literacy, analytical thinking, and teamwork are also seeing a 5–10% wage increase (Mäkelä and Stephany 2024). According to US job vacancy data from 2018–2024, AI-related jobs are significantly more likely to include non-monetary benefits as part of the compensation package, including parental leave and remote working options (Stephany, Mira, and Bone 2025).

Beyond the benefit of not having to commute and the ability to work from anywhere, long-distance and remote roles are often sought after by job seekers due to faster wage growth. In a study comparing the pay trends of remote versus in-office workers in the same occupation, remote workers experienced 4.4% faster annual wage growth, especially in professional and technical sectors (Pabilonia and Vernon 2025). Furthermore, workers who transitioned into remote roles with the same employer saw up to 16 percentage points higher wage growth than their counterparts who remained local (Romem 2024). This demonstrates that switching into remote work can lead to significantly higher wage growth, even within the same company and/or job category.

# 3. Methodology

Below are the steps we took to prepare the dataset to ensure accuracy and statistical integrity in our exploratory data analysis (EDA).

## 3.1 Remove Unnecessary Columns

Firstly, columns containing redundant and irrelevant information were excluded from the dataset. Since the scope of this analysis is focused on job market trends in 2024, it is best practice to remove any outdated NAICS/SOC fields to prevent confusion and duplication. Similarly, metadata fields or duplicate fields that could introduce ambiguity and do not add any meaningful to downstream analysis are excluded. To summarize, unnecessary columns containing the following information were dropped:

* Meta/tracking
* Duplicated location info
* Raw/duplicate title & body
* Duplicated employment info
* Education code columns
* Redundant NAICS/SOC versions
* LOT/V6 occupation hierarchy
* ONET & CIP codes
* Sectors

## 3.2 Handling of missing values

We addressed missing values based on field type and the amount of data missing per field:

### 3.2.1 Numerical Fields

Missing values in key numerical fields were imputed with the median of each respective fields. Since these numerical fields are key for grouping and visualizing trends, simply removing empty rows could reduce the diversity of the dataset and introduce biases. The rationale for imputing the median rather than the mean is that the latter tends to be influenced by outliers and skewed distributions, which is often exhibited in fields like “SALARY. Based on the downstream EDA, additional filtering may be applied to exclude imputed values altogether to prevent distorition and ensure statistical integrity in the trends observed.

### 3.2.2 Categorical Fields

Missing values in key categorical fields were imputed with the placeholder value “Unknown” in order to prevent those rows of data being dropped, leading to unnecessary data loss. By imputing a neutral label like “Unknown”, this strategy ensures that data integrity is retained without introducing false assumptions and biases into the downstream analysis.

### 3.2.3 Columns containing Majority Missing Data

While the above addressed the rationale for imputing missing values, having to impute the majority of a column’s values can also create noise, which provides insignificant information and analytical value to the downstream analysis. Therefore, columns containing more than 50% missing values were excluded from the dataset.

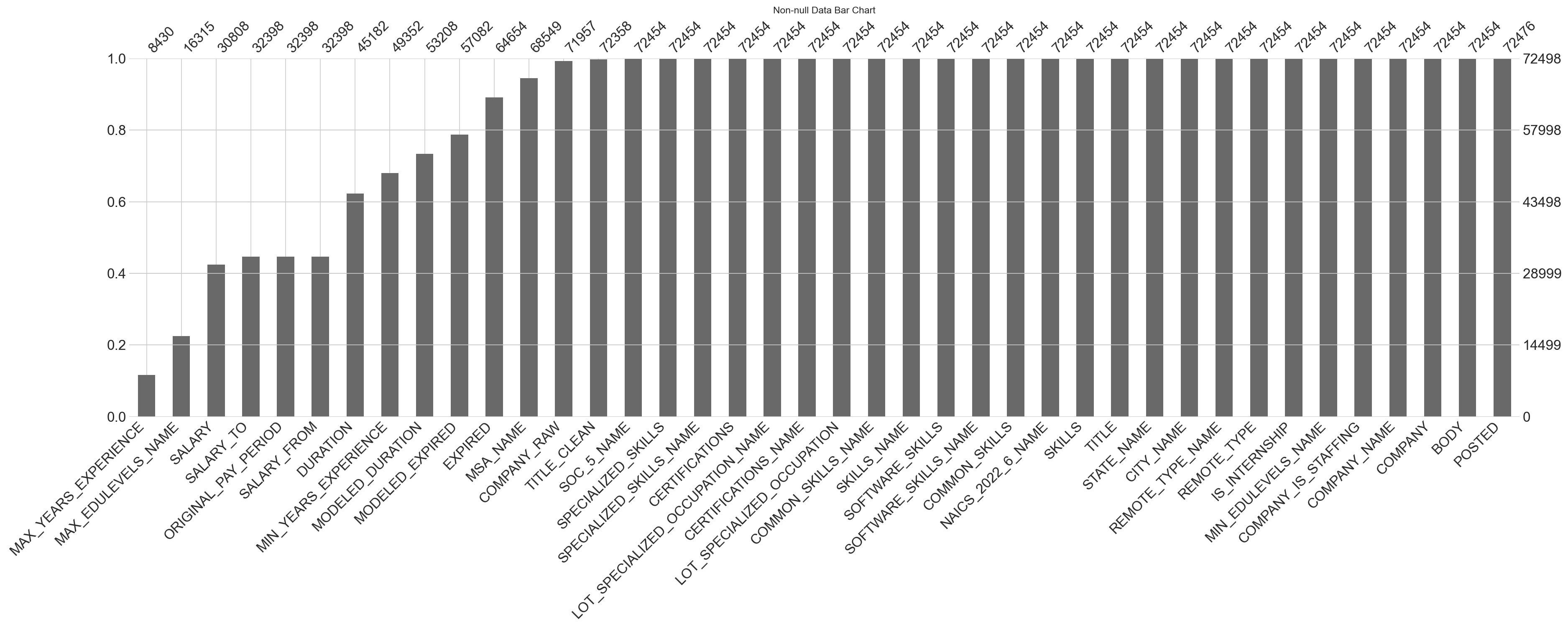


Figure 1: Non-null Data

Column Missing %  
 MAX\_YEARS\_EXPERIENCE 88.372093  
 MAX\_EDULEVELS\_NAME 77.495931  
 SALARY 57.505035  
 SALARY\_TO 55.311871  
 ORIGINAL\_PAY\_PERIOD 55.311871  
 SALARY\_FROM 55.311871  
 DURATION 37.678281  
 MIN\_YEARS\_EXPERIENCE 31.926398  
 MODELED\_DURATION 26.607631  
 MODELED\_EXPIRED 21.264035  
 EXPIRED 10.819609  
 MSA\_NAME 5.447047  
 COMPANY\_RAW 0.746227  
 TITLE\_CLEAN 0.193109  
 SOC\_5\_NAME 0.060691  
 SPECIALIZED\_SKILLS 0.060691  
 SPECIALIZED\_SKILLS\_NAME 0.060691  
 CERTIFICATIONS 0.060691  
LOT\_SPECIALIZED\_OCCUPATION\_NAME 0.060691  
 CERTIFICATIONS\_NAME 0.060691  
 LOT\_SPECIALIZED\_OCCUPATION 0.060691  
 COMMON\_SKILLS\_NAME 0.060691  
 SKILLS\_NAME 0.060691  
 SOFTWARE\_SKILLS 0.060691  
 SOFTWARE\_SKILLS\_NAME 0.060691  
 COMMON\_SKILLS 0.060691  
 NAICS\_2022\_6\_NAME 0.060691  
 SKILLS 0.060691  
 TITLE 0.060691  
 STATE\_NAME 0.060691  
 CITY\_NAME 0.060691  
 REMOTE\_TYPE\_NAME 0.060691  
 REMOTE\_TYPE 0.060691  
 IS\_INTERNSHIP 0.060691  
 MIN\_EDULEVELS\_NAME 0.060691  
 COMPANY\_IS\_STAFFING 0.060691  
 COMPANY\_NAME 0.060691  
 COMPANY 0.060691  
 BODY 0.060691  
 POSTED 0.030346

## 3.3 Remove Duplicates

To eliminate true duplicates from the dataset, job listings that have identical values in all the fields listed below were removed to prevent distortion and ensure statistical integrity:

* Job Title
* Company Name
* Location
* Posting Date
* Skill requirements
* Employment type

# 4. Skill Gap Analysis

## 4.1 Internal Skills Assessment

To evaluate the current technical skill set for each team member, each team member performed a self-rating for a set of core competencies for data analytics and data science, which included Python, SQL, Power BI, Tableau, Excel, Machine Learning, Natural Language Processing (NLP), Cloud Computing, and AWS. The assessment was made on a five-point scale, and the rating values were then compiled into a dataframe for further analysis. We then leveraged a Seaborn heatmap to identify where our strengths concentrated and pinpoint where the gaps existed.

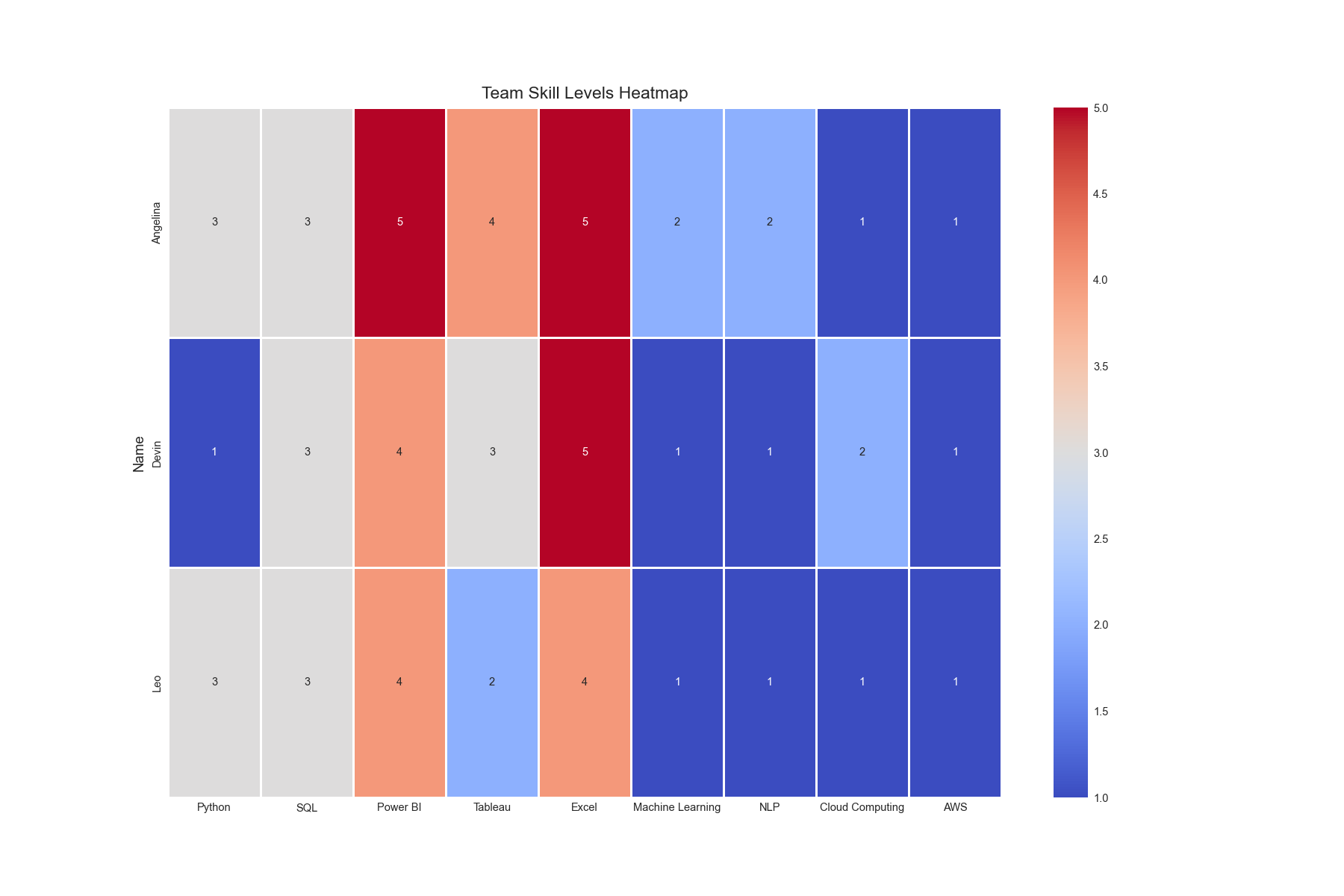


Figure 8: Team Skill Levels Heatmap

**Which skills should each member prioritize learning?**

* **Angelina** – Strong in visualization tools (Power BI 5, Tableau 4, Excel 5). Next priorities: Cloud, AWS, Machine Learning, and NLP. These are high-demand in job postings (Cloud: 64k+, AWS: 10k+, ML/NLP combined: 23k+).
* **Devin** – Solid in Excel (5) and Tableau (3), but weakest in Python, ML, NLP, and AWS. Needs to raise Cloud as well.
* **Leo** – Stronger in Power BI (4) and Excel (4), but very low in ML, NLP, Cloud, and AWS. Should also build up Python and SQL to meet market demand (Python: 17k+, SQL: 43k+ mentions).

## 4.2 External Skills Assessment using Natural Language Processing techniques

To identify the most in-demand skills in the analytics job market, we analyzed the job descriptions from the Lightcast dataset by leveraging Natural Language Processing (NLP) techniques to process the BODY column which contained detailed job descriptions. Below is a summary of the data preparation and extraction process:

* **Text normalization:** To ensure consistency, job descriptions were converted to lowercase, Unicode-normalized format, and whitespaces were stripped.
* **Tokenization:** Numbers and punctuation were filtered out to ensure words can be captured and extracted properly.
* **Stopword removal:** To further filter and concentrate only on meaningful technical words, common English stopwords were removed using Scikit-learn’s built-in stopword list.
* **Keyword filtering:** A predefined list of relevant analytics skills (e.g., Python, SQL, AWS, Tableau, Excel, Pandas, Spark, Machine Learning, NLP, Cloud Computing) was used to identify and count occurrences within the job descriptions.
* **Frequency analysis:** Using Python’s Counter() function, we tallied the frequency of each skill keyword and visualized in a column chart to understand their significance in the job market.

Scanning job descriptions and counting tokens (streaming)...  
Top data analytics skills from job descriptions  
cloud:42787  
sql:35871  
power:21894  
excel:19874  
learning:16361  
tableau:14609  
python:13402  
aws:8510  
machine:5592  
computing:2588  
spark:1496  
docker:613  
pandas:378  
nlp:293  
numpy:187

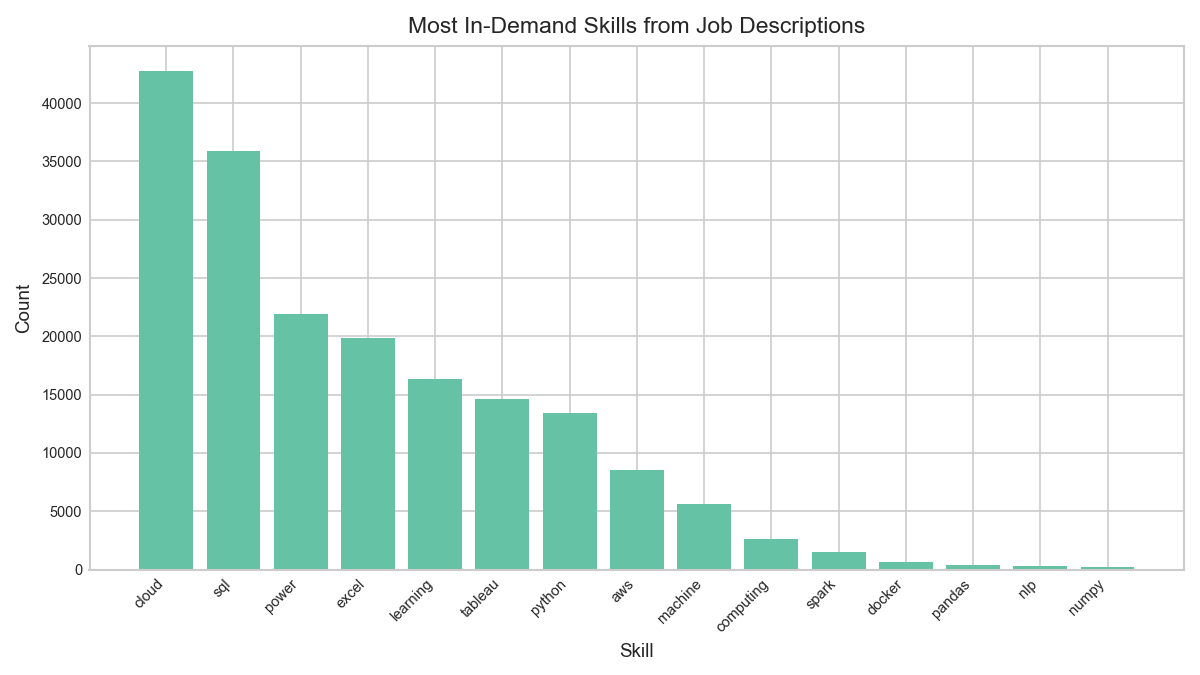


Figure 9: Most In-demand skills from Job Descriptions

## 4.3 Propose an Improvement Plan

According to our analysis, job postings show high demand for Cloud, SQL, Python, ML, and AWS. Our team is strong in visualization (Excel, Power BI, Tableau) but weaker in Cloud, ML, and NLP. This plan aligns our learning with market needs, provides specific resources, and ensures collaboration strategies so the whole team can close the gap together.

**What courses or resources can help?**

* **Cloud & AWS** – free cloud provider tutorials, AWS Educate, and cloud labs for hands-on practice.
* **Machine Learning & NLP** – scikit-learn tutorials, Kaggle competitions, and university modules on ML/NLP.
* **Python & SQL** – interactive platforms (Jupyter notebooks, LeetCode SQL), and official documentation.
* **Docker & Spark** – short online workshops, Spark quickstarts, and Docker “getting started” labs.

**How can the team collaborate to bridge skill gaps?**

* **Role rotation:** assign rotating leads (“cloud lead,” “ML/NLP lead,” “Python/SQL lead”) for mini-projects so each teammate practices outside their strengths.
* **Lightning talks:** weekly 15-minute sessions where one teammate teaches a concept or tool they just learned.
* **Pair programming:** match stronger members (for example, Angelina for visualization) with weaker ones (for example, Devin on Python) to share knowledge in real time.
* **Shared resources:** maintain a team wiki with reusable queries, cloud setup notes, and code snippets.

# 5. Regression, Classification, and Topic Insights

## 5.1 KMeans Clustering Analysis

We performed KMeans clustering on job postings using core features (salary, minimum and maximum years of experience). This analysis seeks to segment jobs into groups with similar compensation and experience profiles, and to interpret these clusters using industry categories (NAICS).

### 5.1.1 Fit KMeans and Assign Clusters - Data Prep

We used KMeans clustering to segment jobs into five groups, using standardized salary and experience as inputs. Each posting is assigned to a cluster.

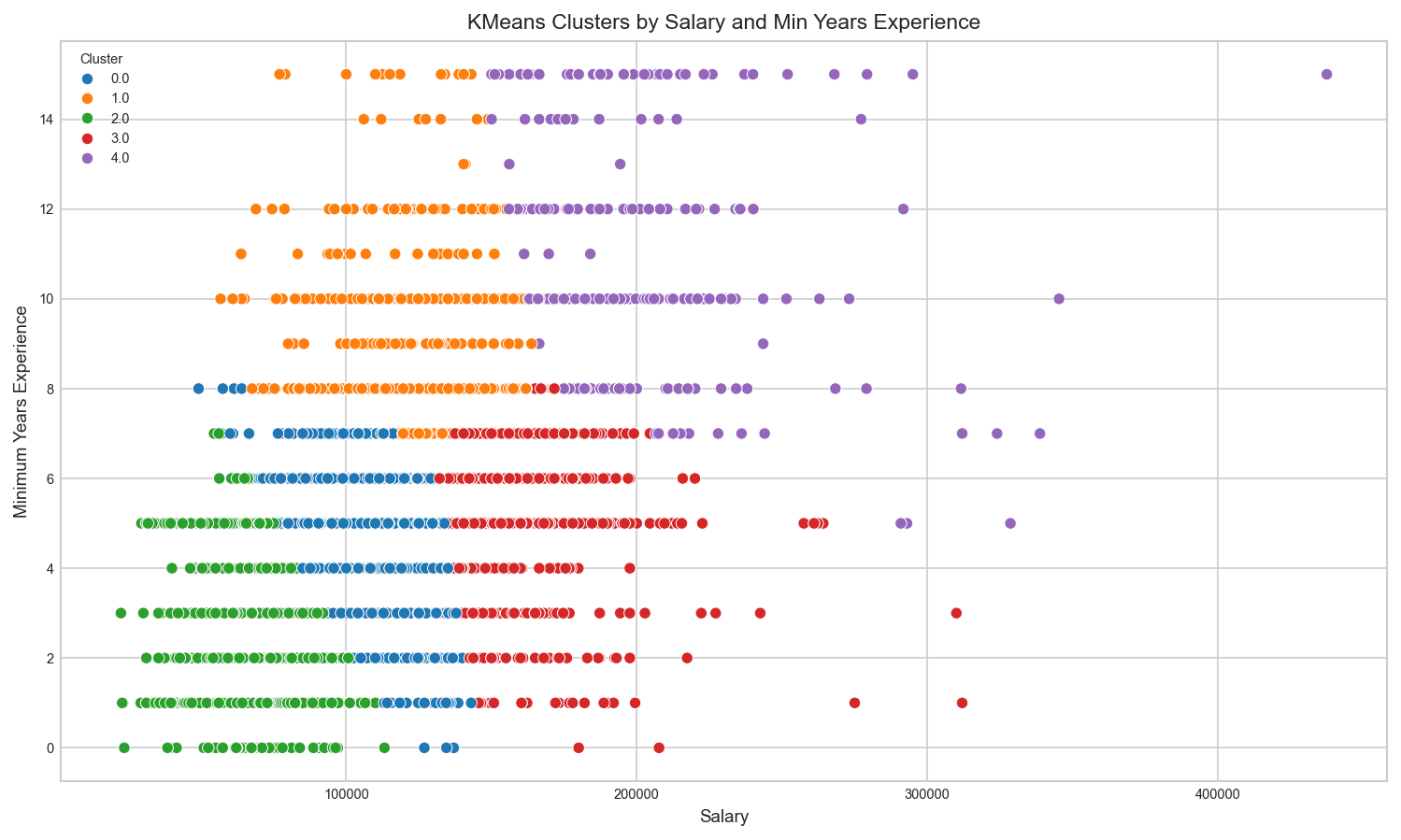


Figure 10: KMeans Clusters by Salary and Min Years Experience

KMeans labels column ready: True  
Unique clusters: 5  
Top 5 Job Titles for Each Cluster:  
  
  
Cluster 0:  
TITLE\_CLEAN  
data analyst 393  
senior data analyst 94  
business intelligence analyst 88  
sr data analyst 37  
data analyst iii 26  
Name: count, dtype: int64  
  
Cluster 1:  
TITLE\_CLEAN  
data analyst 81  
enterprise architect 54  
senior data analyst 23  
solution architect 18  
data modeler 17  
Name: count, dtype: int64  
  
Cluster 2:  
TITLE\_CLEAN  
data analyst 681  
business intelligence analyst 106  
data analyst ii 42  
senior data analyst 35  
research data analyst 26  
Name: count, dtype: int64  
  
Cluster 3:  
TITLE\_CLEAN  
data analyst 68  
enterprise architect 59  
senior data analyst 44  
data engineer analytics 30  
oracle cloud supply chain management senior consultant 19  
Name: count, dtype: int64  
  
Cluster 4:  
TITLE\_CLEAN  
enterprise architect 56  
solution architect 18  
oracle cloud supply chain management manager 14  
principal enterprise architect 14  
enterprise platform architect 12  
Name: count, dtype: int64

### 5.1.2 K-Means Clustering Summary

Clustering on salary and minimum experience produced three clear market segments:

* Entry-level / lower salary
* Mid-career / moderate salary
* Senior specialist / high salary

This segmentation aligns with career ladders and confirms the broad, positive relationship between experience and compensation.

**Business relevance**

* **Job seekers**: Use cluster patterns to target industries/roles occupying higher-pay segments and to plan upskilling.
* **Employers**: Map openings to cluster ranges to calibrate pay bands against the external market and reduce attrition risk.

## 5.2 Regression – Predicting Salary

We chose Random Forest over Linear Regression because the relationship between experience and salary is non-linear, and the model captures complex interactions more effectively.

The goal of this model is to predict job posting salary using experience and employment type features. The model uses an 80/20 train–test split and evaluates performance using RMSE and R² metrics.

Root Mean Squared Error (RMSE): 34744.38  
R² Score: 0.336

### 5.2.1 Regression Results and Interpretation

We modeled salary using Minimum Years of Experience (MIN\_YEARS\_EXPERIENCE) and Remote Type (REMOTE\_TYPE\_NAME). We excluded MAX\_YEARS\_EXPERIENCE (>85% missing) to protect statistical integrity; this reduced sample size but improved reliability.

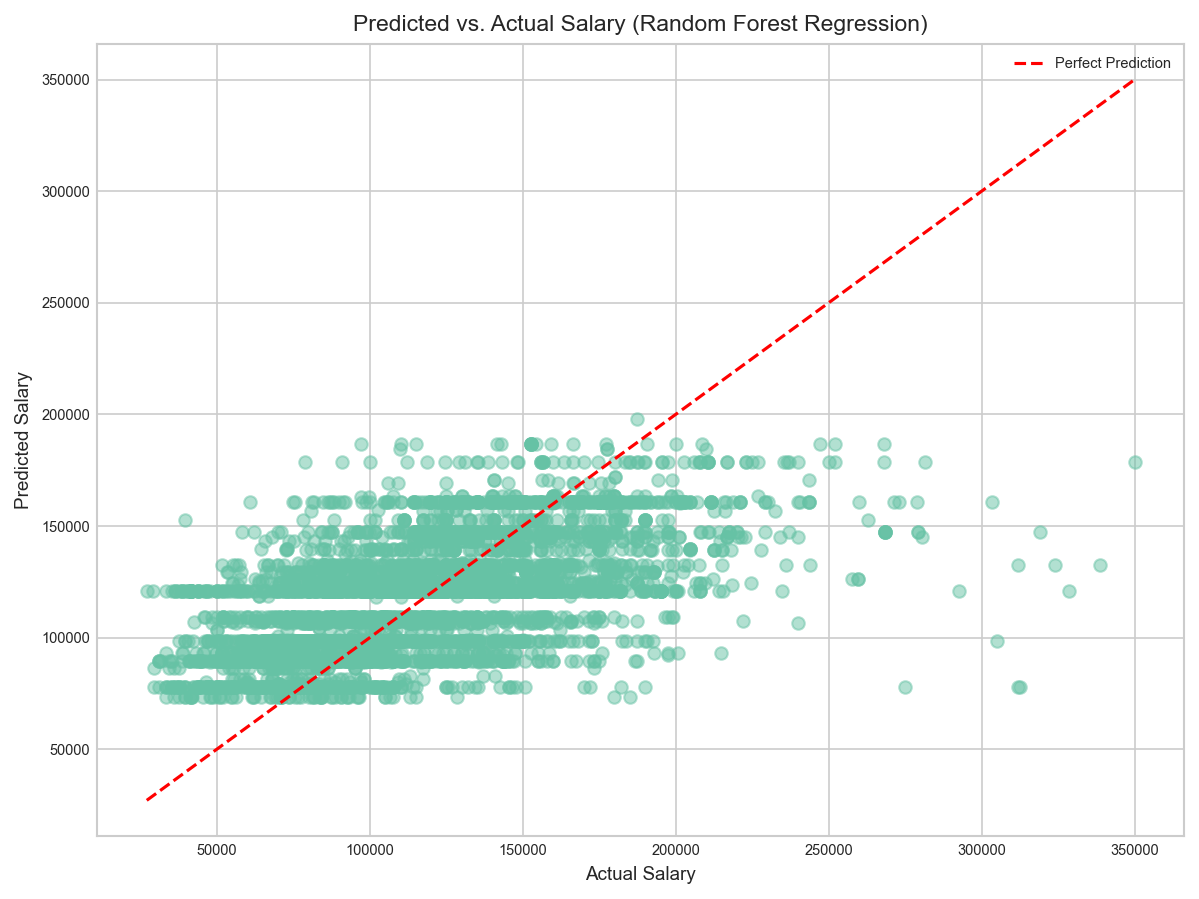
**Model performance**

* **RMSE**: 34,744.38
* **R²**: 0.336

**What this means** An R² of 0.336 indicates experience and remote status explain a meaningful, but incomplete share of salary variation. That’s consistent with cross-industry labor data, where compensation is also driven by occupation, industry, location, and scarce technical capabilities.

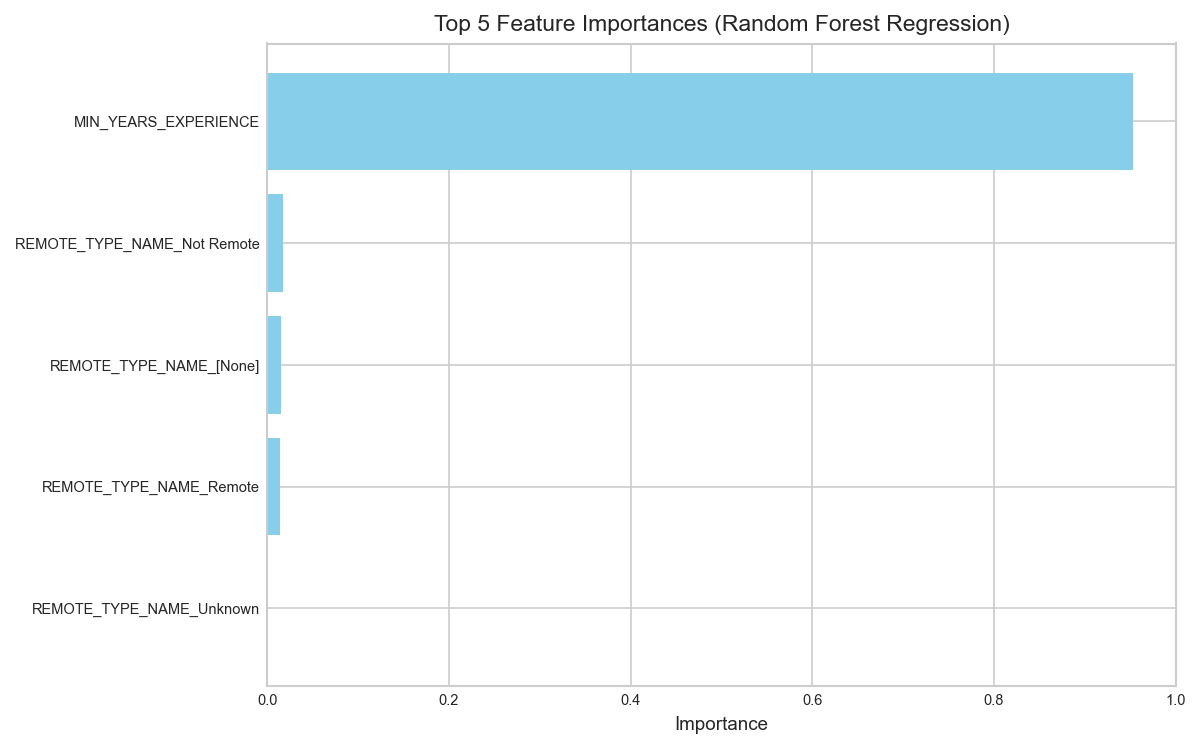
**Business relevance**

* **For job seekers**: Experience contributes to higher earnings, but targeted specialization (e.g., analytics, cloud, data engineering) and industry selection are decisive for larger pay jumps.
* **For employers**: Pricing talent solely by tenure can misalign offers in high-skill roles. Clear remote/on-site definitions in postings improve candidate signal and reduce noise in market benchmarking.



Predicted vs. Actual Salary (Random Forest Regression)

Top 5 features influencing salary prediction:  
MIN\_YEARS\_EXPERIENCE: 0.953  
REMOTE\_TYPE\_NAME\_Not Remote: 0.017  
REMOTE\_TYPE\_NAME\_[None]: 0.015  
REMOTE\_TYPE\_NAME\_Remote: 0.014  
REMOTE\_TYPE\_NAME\_Unknown: 0.000



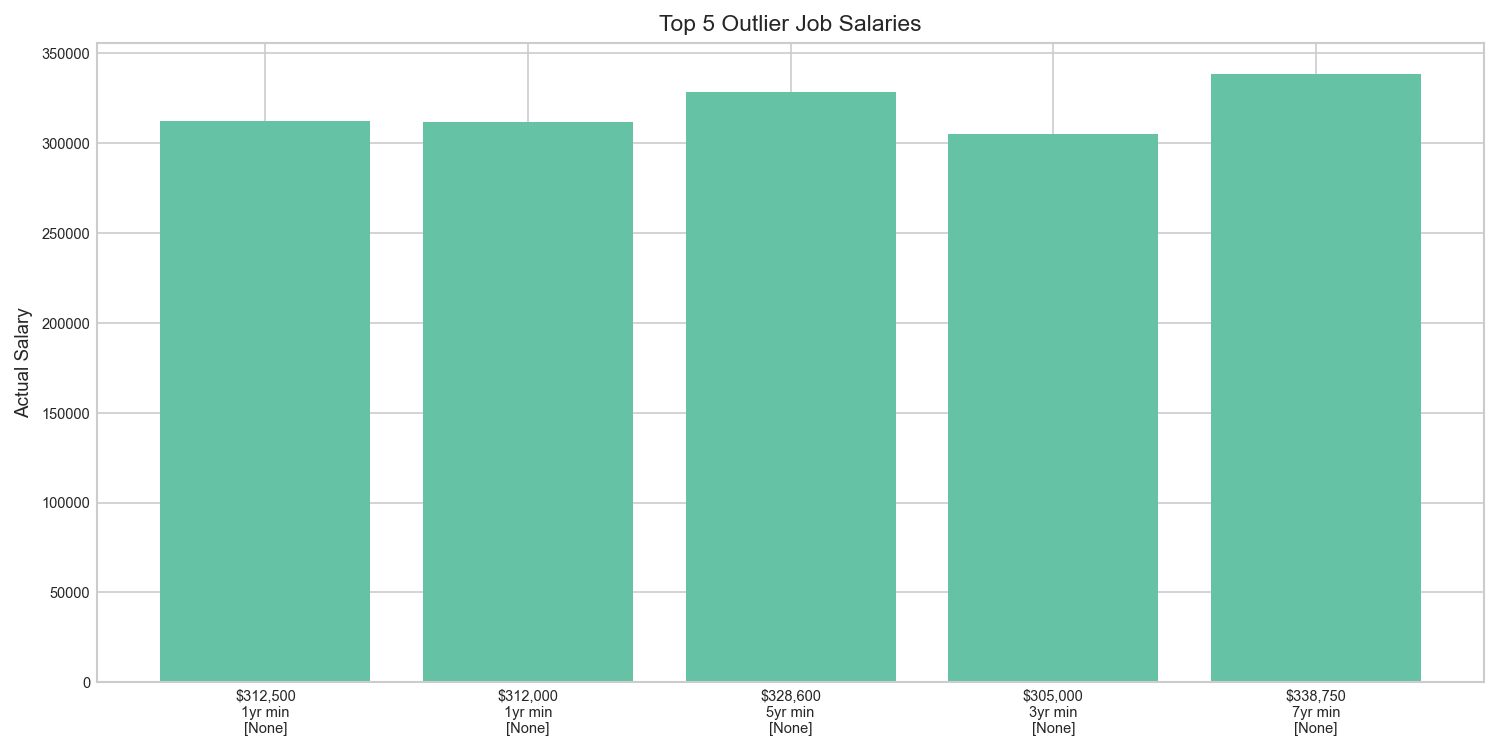
Top 5 Feature Importances (Random Forest Regression)

SALARY MIN\_YEARS\_EXPERIENCE REMOTE\_TYPE\_NAME  
10189 312500.0 1.0 [None]  
19209 312000.0 1.0 [None]  
31552 328600.0 5.0 [None]  
2258 305000.0 3.0 [None]  
29554 338750.0 7.0 [None]

### 5.2.2 Outlier Jobs and Market Signals

The largest residuals include postings above $300,000 with only 1–5 years minimum experience, often missing a remote/on-site label. These cases point to specialized, high-impact roles where compensation is decoupled from tenure (e.g., advanced analytics, cloud/platform, niche leadership).

A small number of postings show very high pay at low experience with unspecified remote type. These likely reflect specialized roles or incomplete records. In practice, validate or isolate these cases before setting salary bands or training production models.



Top 5 Outlier Job Salaries

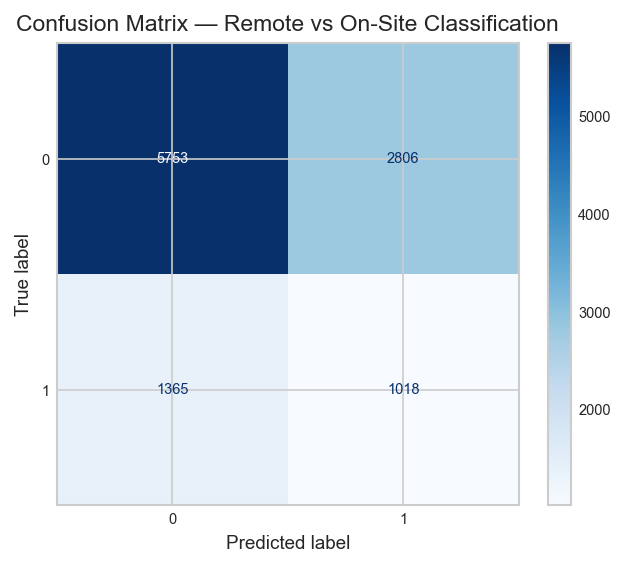
**Business relevance**

* **Employers**: Standardize and QA high-pay, low-tenure postings, unclear fields and atypical mixes should be reviewed before publication.
* **Job seekers**: Outliers highlight skill paths where focused upskilling can command premium pay earlier in a career.

## 5.3 Classification – Predicting Remote vs On-Site Job

We trained a logistic regression to predict Remote vs On-Site using MIN\_YEARS\_EXPERIENCE plus an available categorical descriptor (e.g., employment type or state).

Accuracy: 0.619, F1 Score: 0.328



Confusion Matrix — Remote vs On-Site Classification

### 5.3.1 Classification Results and Interpretation

**Performance:**

* **Accuracy**: 0.622
* **F1 Score**: 0.327

**What this means:** The model distinguishes remote vs. on-site at a moderate level, consistent with the idea that remote status is policy/role-design driven, not primarily a function of required experience.

**Business relevance:**

* **Employers:** Remote flexibility can be offered across experience levels without materially distorting supply. If remote status is strategic, emphasize role/industry signals rather than tenure in postings.
* **Job seekers:** Remote options exist from entry to senior levels, broadening geographic reach and negotiation leverage.

# 6. Conclusion & Recommendation

This project, which analyzed the 2024 job market, provided key insights into the factors influencing compensation and future workforce development. Our analysis confirms that the employment landscape is being fundamentally reshaped by technology and the demand for flexible work, creating both challenges and opportunities for job seekers and employers.

## 6.1 Key Findings and Market Dynamics

Skill-Driven Wage Premiums Our Natural Language Processing (NLP) analysis identified a clear difference between the team’s strengths and current market demand. While the team performs well in visualization tools such as Excel, Power BI, and Tableau, the job market places greater emphasis on Cloud Computing, SQL, Python, and Machine Learning (ML). The frequency of these skills in job postings shows that focused development in these areas supports higher compensation and continued competitiveness in analytics roles. Experience as the Primary Salary Driver The Random Forest Regression model achieved an RMSE of 34,744.38 and an R² of 0.336, identifying Minimum Years of Experience as the most influential feature in predicting salary, with a feature importance of approximately 0.953. This result indicates that, while technical specialization enhances access to well-paid roles, accumulated experience continues to be the main factor influencing salary progression. The Segmented Job Market The KMeans Clustering model grouped job postings into five clusters, summarized into three main tiers—Entry-Level (lower salary), Mid-Career (moderate salary), and Senior Specialist (higher salary). This pattern aligns with traditional career development paths and suggests that job seekers should aim for higher-tier roles by combining experience with proficiency in high-demand skills such as Cloud Computing, Python, and Machine Learning. Remote Work Status is Not Experience-Dependent The Logistic Regression model achieved an accuracy of 0.622 and an F1 score of 0.327, showing moderate predictive strength when using experience to classify remote versus on-site jobs. This suggests that remote flexibility depends more on company policy or role design than on years of experience. Remote opportunities appear across all experience tiers, providing both entry-level and senior professionals with broader access to roles.

## 6.2 Strategic Recommendations

Based on these findings, the following strategies are recommended for both job seekers (the team) and employers: Continue Upskilling in Cloud and Machine Learning Efforts should focus on further developing Cloud/AWS and Machine Learning/NLP skills. These areas show the greatest opportunity for alignment with market needs and can improve access to higher-tier roles. Standardize High-Value Job Postings Employers should review and standardize postings with high salaries and lower experience requirements. Accurate labeling of REMOTE\_TYPE and clear definitions of required technical skills will improve market consistency, strengthen salary benchmarking, and attract well-matched candidates. Use Remote Flexibility to Expand Talent Pools Since remote status does not strongly depend on experience, employers can apply flexible work policies to reach qualified candidates at all career levels. This approach supports broader hiring while maintaining consistency in pay structures.

Mäkelä, E., and F. Stephany. 2024: “[Complement or substitute? How AI increases the demand for human skills](https://arxiv.org/abs/2412.19754),” *arXiv preprint arXiv:2412.19754*,.

Pabilonia, S. W., and V. Vernon. 2025: “[Remote work, wages, and hours worked in the united states](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4561618),” *SSRN Electronic Journal*,.

PwC. 2024: “[PwC 2024 Global AI Jobs Barometer: Demand, Wages, and Disruption](https://www.pwc.com/gx/en/news-room/press-releases/2024/pwc-2024-global-ai-jobs-barometer.html),”

Romem, I. 2024: “[Long-Distance Work and Compensation](https://www.adpresearch.com/long-distance-work-and-compensation/),”ADP Research Institute.

Stephany, F., A. Mira, and M. Bone. 2025: “[Beyond pay: AI skills reward more job benefits](https://arxiv.org/abs/2507.20410),” *arXiv preprint arXiv:2507.20410*,.