Toronto Metropolitan University
Data Science & Analytics Graduate Studies

Literature Review and Exploratory Data Analysis

ResumeMatch: A Dual-Role AI-Driven Platform for Transparent Resume-Job Compatibility and Enhanced Hiring Outcomes

Rabia Danish
Supervisor's Name:
Professor Dr. Pawel Pralat
Date of submission:
24th June 2025

Table of Contents

Abstract

This research project proposes ResumeMatch, a dual-role AI-driven platform designed to improve resume-job compatibility analysis through advanced natural language processing (NLP) and large language models (LLMs). Existing recruitment systems often rely on rigid keyword-based filtering, resulting in mismatched candidate evaluations and missed hiring opportunities. Job seekers face rejection due to formatting or keyword issues, while recruiters struggle to identify transferable skills among applicants. Leveraging embedding models for semantic similarity scoring and LLMs for content generation, ResumeMatch offers transparent, personalized insights for both user types. For job seekers, the system analyzes resumes against job descriptions, visualizes matched and missing skills, provides actionable improvement tips, and generates customized interview questions. For recruiters, it ranks candidates based on relevance, breaks down compatibility by skill and experience, and presents interactive dashboards to support decision-making. Using publicly available resume and job posting datasets from Kaggle, the platform aims to bridge key gaps in contextual understanding, fairness, and usability, ultimately enhancing the effectiveness and transparency of AI-assisted hiring processes.

Introduction

The integration of artificial intelligence (AI) into recruitment and human resources (HR) has profoundly transformed talent acquisition. Advances in machine learning (ML), natural language processing (NLP), and deep learning have enabled AI systems to screen candidates, predict job fit, and offer real-time feedback. These technologies promise increased efficiency, scalability, and the potential to reduce human biases in hiring. However, significant challenges persist, particularly concerning fairness, interpretability, generalizability, and ethical considerations. This review synthesizes sixteen peer-reviewed studies on AI-driven recruitment, covering resume analysis, bias detection, job recommendation systems, and fairness-aware algorithm development, to provide a comprehensive understanding of the current state and challenges.

Literature Review

Recent literature reflects a growing interest in AI-based resume evaluation systems that leverage advanced natural language processing and machine learning algorithms. The study conducted by Sruthi et al. (2023) introduced a smart resume analyzer using recurrent neural networks (RNNs) to extract key information and suggest improvements for resume quality and skill development. While effective in its domain, its limited scope to computer science resumes highlights a common issue: domain-specificity reduces generalizability.

Addressing scalability, Bhatt et al. (2024) applied MapReduce techniques for large-scale resume dataset preprocessing, significantly improving the efficiency of downstream ML algorithms like K-Nearest Neighbors. This method achieved an impressive 97.2% F1-score, demonstrating the potential of combining big data processing with classification algorithms. Nevertheless, the authors noted the lack of standardized benchmarks and varied performance across diverse resume structures and industries.

Several studies highlight the benefits of deep learning architectures for information extraction. Ayishathahira et al. (2018) implemented a system combining CNN and Bi-LSTM-CRF models to parse 23 distinct fields from resumes. Despite a relatively small dataset (approximately 1200 resumes), their work showed deep learning's superiority over traditional rule-based parsing. However, format variability and file-type inconsistencies remained significant obstacles. More recent efforts by Kinger et al. (2024) integrated YOLOv5 for section identification and DistilBERT for entity recognition, achieving a parsing accuracy of 96.2%. This pipeline improved Applicant Tracking Systems (ATS) by enabling fine-grained classification and contextual extraction. Still, challenges in handling domain variation and language diversity limit widespread deployment, especially across multilingual platforms.

The increasing sophistication of embeddings has led to innovative recommender systems. Bevara et al. (2025) developed Resume2Vec, using transformer-generated embeddings to align candidate resumes with job descriptions more accurately than traditional keyword-based ATS. This method showed substantial improvements in ranking accuracy using Normalized Discounted Cumulative Gain (nDCG) and Ranked Biased Overlap (RBO), particularly in mechanical engineering and healthcare. While human evaluation enhanced credibility, the use of student raters instead of HR professionals weakened external validity. Furthermore, interpretability and computational efficiency remain unresolved for production-level deployment.

Fairness and ethical concerns are recurring motifs in AI-based hiring research. Delecraz et al. (2022) developed a job matching algorithm specifically for NEET (Not in Employment, Education, or Training) populations, incorporating fairness metrics like Disparate Impact and Statistical Parity. Their study exposed systemic underrepresentation in job recommendations, particularly for foreign nationals and applicants requiring work permits. While the model achieved parity for gender, persistent disparities for attributes like nationality suggest that algorithmic fairness requires deeper structural changes in data collection and organizational policy.

Similar concerns were explored in a study conducted by Frissen et al. (2023), who proposed a system to classify job advertisements into five categories of discriminatory language using Random Forest classifiers and FastText embeddings. Although effective in identifying exclusionary language, its scope was limited to job descriptions. The authors recommended integrating such systems into hiring pipelines to holistically evaluate diversity and inclusivity, including the creation of fairness scores.

Beyond technical implementations, public perception of AI hiring tools is a critical yet understudied area. Zhang and Yencha (2022) found general skepticism among users, especially towards video interview screening tools. Resume-screening algorithms were marginally more acceptable, but fairness and effectiveness were questioned across demographic groups. Higher income and education levels correlated with more favorable perceptions, underscoring the need for user-centered design alongside algorithmic improvements.

The ethical implications of AI in sensitive sectors, such as healthcare, were examined by Burrell and Mcandrew (2023). Their consultant-based intervention highlighted the need for continuous audits, inclusive datasets, and cultural change to combat discrimination. This research emphasizes that AI must

be embedded within organizational structures that prioritize ethical hiring, calling attention to policy development and systemic accountability rather than just technical solutions.

In the context of recommender systems for large-scale deployments, Zhao et al. (2021) implemented a two-stage embedding-based system using fused embeddings and contextual reranking. This system yielded a 104% increase in click-through rate and a 37% improvement in nDCG over baselines, validating deep representation learning in production environments. However, concerns persist around computational load and reproducibility due to proprietary data and algorithms.

Other recommender approaches include CCRS, a reciprocal matching model by Özcan & Ögüdücü (2017) that accounts for both candidate and employer preferences. By incorporating TrustRank and user interaction features, CCRS addressed cold start and data sparsity, outperforming baseline classifiers. However, the limited dataset (7455 applications) and lack of demographic diversity raise questions about the model's robustness in varied recruitment scenarios.

Studies also investigated named entity recognition (NER) for specific resume components. Gaur et al. (2021) developed a semi-supervised deep learning approach to identify educational institutions and degrees with minimal labeled data, achieving 92.06% accuracy. This technique addressed the challenge of scarce annotated resume datasets, though it remained constrained to the education section. The absence of a standard dataset for NER tasks in resumes continues to hinder benchmarking.

Another important line of research involves forecasting labor market trends. Fettach et al. (2025) used temporal knowledge graph embeddings to predict skill demand in the Moroccan IT sector, demonstrating the feasibility of time-aware link prediction for anticipating emerging job skill requirements. Limitations included the lack of integration with real-time labor data and difficulty applying the model in open-world settings.

A few studies combined AI systems with real-time user interaction features. Singh and Gupta (2023) designed a recommendation engine that incorporated feedback loops, skill ontologies, and social boosting mechanisms to dynamically recalibrate candidate rankings. This improved screening speed and precision, but authors noted practical challenges such as organizational reluctance to invest in AI and the need for more transparent algorithmic decision-making.

Building on real-time systems, J et al. (2024) proposed a CNN-based resume analyzer that predicted suitable job roles with 99.48% accuracy and generated feedback and course suggestions based on resume-job description compatibility. While outperforming traditional ML algorithms, limitations like restricted language support and oversimplified resume scoring metrics point to gaps in inclusivity and nuance. The authors recommended increasing dataset diversity and adopting more sophisticated pattern-matching techniques, reinforcing the need for richer, more representative training data.

Other systems also aim to match resumes with jobs while providing applicant feedback. Pabalkar et al. (2024) presented an automated screening system combining TF-IDF and nine classifiers to match resumes to job descriptions, providing scoring and recommendations. Their Linear SVM model achieved 78.53% accuracy but struggled with precision compared to deep learning counterparts. The

authors advocated for integrating with social media profiles (e.g., GitHub, LinkedIn) to improve context and personalization, aligning with growing interest in multi-modal recruitment systems.

Limitations and Future Directions

While the literature demonstrates meaningful progress in AI-driven recruitment, several persistent limitations remain. A major concern is the lack of standardized, large-scale, and publicly available datasets. Many studies rely on small or proprietary datasets (often fewer than 15,000 resumes), which limits generalizability and hinders reproducibility across research. This restricts the applicability of trained models in real-world recruitment settings and challenges efforts to benchmark system performance consistently (Gaur et al., 2021).

Another gap lies in the disconnect between algorithmic predictions and real-world hiring outcomes. While many models report high accuracy or ranking performance, few are validated using longitudinal or field data such as actual hiring decisions or post-hire success. This lack of external validation weakens confidence in system reliability and limits practical adoption. Zhao et al. (2021) made progress by integrating user interaction metrics, but comprehensive industry validation remains scarce.

Most systems are also limited in their language and cultural scope. The majority of models are trained on English-language data, making them unsuitable for global or multilingual recruitment use cases. Additionally, fairness interventions tend to focus on individual demographic features, such as gender or nationality without addressing intersectionality or socio-economic factors that influence hiring bias (Bevara et al., 2025; Delecraz et al., 2022).

Explainability remains a significant challenge. Many AI models function as opaque black boxes, offering users little insight into how recommendations are generated. In the context of employment, this lack of transparency undermines trust and accountability. While a few studies suggest integrating visualization or interpretability tools, most fail to implement user-centered design practices. Our research addresses this by providing clear skill match breakdowns, gap visualizations, and interpretable scoring metrics for both job seekers and recruiters.

Ethical and legal considerations are also underexplored. Few models directly address compliance with frameworks like GDPR or EEOC guidelines. Studies by Burrell and Mcandrew (2023) and Frissen et al. (2023) underscore the need for continuous audits and fairness evaluations, but there remains a lack of systematic integration of these safeguards into technical systems.

Finally, static models cannot keep pace with evolving labor demands. Most tools do not incorporate dynamic learning or real-time market adaptation. Fettach et al. (2025) proposed temporal embeddings for skill forecasting, highlighting the potential of adaptive systems. Building on this, our research is conducted on future integration with real-time job market data and dynamic skill recommendation modules powered by LLMs.

Conclusion

AI has the potential to enhance recruitment through efficient, scalable, and data-driven methods, yet existing solutions are limited by small datasets, narrow cultural scope, lack of transparency, and

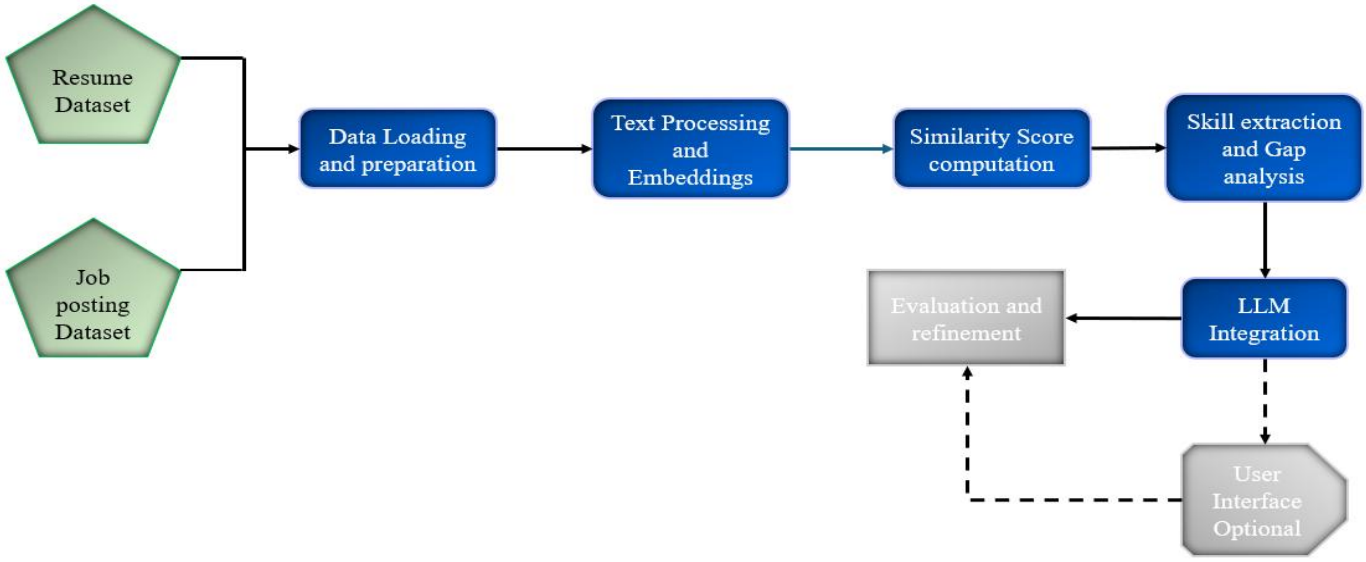
insufficient validation. Our dual-role platform directly responds to these gaps by offering transparent, explainable, and adaptable features for both job seekers and recruiters. Through modular design, dynamic scoring, skill visualization, and integration of LLMs for personalized insights, the system is built to evolve with user needs and labor trends. Moving forward, recruitment technologies must prioritize not just accuracy, but fairness, inclusivity, and accountability to realize the full promise of AI in hiring.

Proposed Tentative Methodology for the Project

This project aims to develop a dual-role AI-powered platform that analyzes and matches resumes with job descriptions using advanced NLP and LLM techniques. The methodology is guided by recent research highlighting the limitations of traditional keyword-based recruitment systems (Sruthi et al., 2023; Delecraz et al., 2022) and advocating for embedding-based, interpretable models (Bevara et al., 2025; Zhao et al., 2021). The tentative methodology consists of the following stages:

1. **Data Loading and Preparation:** Load and clean the Resume and Job Postings datasets.
2. **Text Processing and Embedding:** Process job descriptions and resumes (e.g., extract relevant sections, clean text) and convert them into numerical representations (embeddings) using appropriate models.
3. **Similarity Score Computation:** Compute similarity scores between resumes and job descriptions across multiple dimensions, such as skills, experience, and overall semantic alignment using cosine similarity or other suitable metrics.
4. **Skill Extraction and Gap Analysis:** Identify and compare skills present in resumes against those required in job descriptions. Visualize skill matches and gaps using tools like radar plots or bar charts.
5. **LLM-Generated Insights:** Utilize LLMs to generate content such as suggested resume improvements, relevant interview questions, and detailed breakdowns of matching factors for recruiters.
6. **Platform Development (Optional):** Develop a lightweight interface using Streamlit to enable interaction for both job seekers and recruiters. This may include uploading files, viewing similarity scores, and visualizations.
7. **Evaluation and Refinement:** Conduct evaluations of match accuracy and insight usefulness. Refine embeddings, prompt engineering, or UI components based on findings.

This tentative methodology ensures the system is both technically robust and user-centered. The dual-role design addresses gaps identified in the literature namely, lack of transparency, weak personalization, and limited recruiter-facing insights while offering a foundation for further research and platform expansion.



Data Preparation

The data preparation phase involved collecting, cleaning, and merging disparate datasets to create a comprehensive foundation for analysis. The objective was to consolidate job posting details, company information, salary ranges, and associated skills into a single, cohesive dataset, and to prepare the resume data for matching.

Data Collection

The datasets utilized in this project were sourced from Kaggle Inc. and comprise two main components:

1. **LinkedIn Job Postings Dataset**: This dataset contains a nearly comprehensive record of 124,000+ job postings listed in 2023 and 2024 on LinkedIn. It is composed of several inter-related CSV files:
    - postings.csv: Core job advertisement details (title, description, location, company, work type, salary, etc.).
    - jobs/job_skills.csv: Maps job IDs to abstract skill identifiers.
    - mappings/skills.csv: Provides names for the abstract skill identifiers.
    - jobs/salaries.csv: Detailed salary information, including min, max, median, and pay period.
    - companies/companies.csv: General information about companies (name, size, location).
    - companies/company_industries.csv: Maps companies to their respective industries.
2. **Resume Dataset**: This dataset contains 2400+ Resumes in string as well as PDF format categorized by job role. It consists of:
    - Resume/Resume.csv: Raw resume text and their assigned categories.

Data Dictionary

The dataset can be described briefly as follows.

JOB_POSTINGS

Column	Description	Variable Type
job_id	The job ID from LinkedIn	string
company_id	ID for the company (maps to companies.csv)	string
title	Job title	string
description	Job description	string
max_salary	Maximum salary	float
med_salary	Median salary	float
min_salary	Minimum salary	float
pay_period	Pay period (Hourly, Monthly, Yearly)	string
formatted_work_type	Type of work (Fulltime, Parttime, Contract)	string
location	Job location	string
applies	Number of applications submitted	int
original_listed_time	Original time job was listed	datetime
remote_allowed	Whether job permits remote work	boolean
views	Number of views	int
job_posting_url	URL of the job posting	string
application_url	Application submission URL	string
application_type	Type of application (offsite, onsite)	string
expiry	Expiration date of listing	datetime
closed_time	Time job listing closed	datetime
formatted_experience_level	Experience level (entry, executive, etc.)	string
skills_desc	Required skills description	string
listed_time	Time when job was listed	datetime
posting_domain	Domain of application website	string

sponsored	If the job listing is promoted	boolean
work_type	Type of work	string
currency	Currency of salary	string
compensation_type	Type of compensation	string

SALARIES

Column	Description	Variable Type
salary_id	Unique salary ID	string
job_id	Foreign key from jobs table	string
max_salary	Maximum salary	float
med_salary	Median salary	float
min_salary	Minimum salary	float
pay_period	Pay frequency (Hourly, Monthly, Yearly)	string
currency	Currency of salary	string
compensation_type	Type of compensation (Fixed, Variable, etc)	string

COMPANIES

Column	Description	Variable Type
company_id	LinkedIn Company ID	string
name	Company name	string
description	Company description	string
company_size	Company size group (0 Smallest – 7 Largest)	int
country	Country of HQ	string
state	State of HQ	string
city	City of HQ	string
zip_code	ZIP code of HQ	string
address	Street address of HQ	string

url	LinkedIn URL of company	string

SKILLS

Column	Description	Variable Type
skill_abr	Skill abbreviation	string
skill_name	Full skill name	string

JOB_SKILLS

Column	Description	Variable Type
job_id	Job ID (references jobs table)	string
skill_abr	Skill abbreviation (foreign key)	string

COMPANY_INDUSTRIES

Column	Description	Variable Type
company_id	Company ID (references companies)	string
industry	Industry ID	string

Resume

Column	Description	Variable Type
ID	Unique identifier and file name for the respective pdf	string
Resume_str	Contains the resume text only in string format	string
Resume_html	Contains the resume data in html format	string
Category	Category of the job the resume was used to apply	string

Data Merging and Initial Cleaning

The initial datasets were raw and fragmented. A critical step was to merge them into a single, comprehensive DataFrame (job_postings) for job postings, and to prepare the resume_df. This process involved several key steps:

1. **Skill Aggregation:** The job_skills and skills tables were joined, and then skills were aggregated by job_id into a comma-separated list (job_required_skills_list).
2. **Company Information Consolidation:** companies and company_industries were merged. Company industries were also aggregated by company_id (company_industries_list).
3. **Salary Normalization:** The salaries table, which contained min/max/median salaries across various pay periods (hourly, weekly, biweekly, monthly, yearly), was processed to convert all salaries to a consistent calculated_yearly_salary. This involved hourly/weekly/biweekly/monthly periods to estimate annual equivalents.
4. **Sequential Merging for job_postings:**
    - postings was used as the base.
    - calculated_yearly_salary was merged.
    - Company details (company_name, company_size, state, country, city, address, zip_code) were merged. The company_name from the companies table was prioritized over postings.csv if available.
    - company_industries_list was merged.
    - job_required_skills_list was merged.
    - A final_yearly_salary column was created, preferring normalized_salary from postings and falling back to calculated_yearly_salary from salaries.csv. Redundant salary columns and other columns that does not provide meaningful information were then dropped.

Dataset Summary and Description

The final job_postings DataFrame contains 122,890 job postings with 16 columns, providing details such as job title, company name, location, salary information, work type, experience level, skills, and industry.

	job_id	company_name	job_title	pay_period	job_posting_location	Work_type	remote_allowed	Experience_level	sponsored	currency	compensation_type	company_industries_list	skills_list	final_yearly_salary	job_description	skills_description
0	921716	corcoran sawyer smith	Marketing Coordinator	hourly	Princeton, NJ	full-time	False	Not Specified	0	usd	base_salary	real estate	marketing,sales	38480.0	Job description A leading real estate firm in ...	requirements: we are seeking a college or grad...
1	1829192	Not Specified	Mental Health Therapist/Counselor	hourly	Fort Collins, CO	full-time	False	Not Specified	0	usd	base_salary	Not Specified	health care provider	83200.0	At Aspen Therapy and Wellness , we are committ...	Not Specified
2	10998357	the national exemplar	Assitant Restaurant Manager	yearly	Cincinnati, OH	full-time	False	Not Specified	0	usd	base_salary	restaurants	management,manufacturing	55000.0	The National Exemplar is accepting application...	we are currently accepting resumes for foh - ...
3	23221523	abrams fensterman, llp	Senior Elder Law / Trusts and Estates Associat...	yearly	New Hyde Park, NY	full-time	False	Not Specified	0	usd	base_salary	law practice	other	157500.0	Senior Associate Attorney - Elder Law / Trus...	this position requires a baseline understandin...
4	35982263	Not Specified	Service Technician	yearly	Burlington, IA	full-time	False	Not Specified	0	usd	base_salary	Not Specified	information technology	70000.0	Looking for HVAC service tech with experience ...	Not Specified

The resume_df DataFrame contains 2,482 resumes with 3 columns: ID, Category, cleaned resume text.

	ID	Category	cleaned_resume
0	16852973	HR	HR ADMINISTRATOR/MARKETING ASSOCIATE HR ADMINI...
1	22323967	HR	HR SPECIALIST, US HR OPERATIONS Summary Versat...
2	33176873	HR	HR DIRECTOR Summary Over 20 years experience i...
3	27018550	HR	HR SPECIALIST Summary Dedicated, Driven, and D...
4	17812897	HR	HR MANAGER Skill Highlights HR SKILLS HR Depar...

Preprocessing Steps

After loading and merging the datasets, a thorough check for missing values and duplicate rows was performed. This step is critical for data integrity and to ensure that subsequent analyses are based on complete and unique records.

```
final_job_df.isna().sum().sort_values(ascending=False)
```

	0
skills_description	121410
remote_allowed	108603
applies	100529
compensation_type	87776
final_yearly_salary	87776
currency	87776
pay_period	87776
Experience_level	29409
job_required_skills_list	1753
company_name	1719
company_industries_list	1718
company_id	1717
views	1689
job_description	7
job_id	0
job_title	0
job_posting_location	0
Work_type	0
sponsored	0

- **Duplicate Removal:** Duplicates were removed based on job_id for job postings (no duplicates found) and Resume_str for resumes (2 duplicates removed).
- **Handling Missing Values and Standardization:**
    - Missing values in remote_allowed were filled with 0 and the column was converted to boolean.
    - final_yearly_salary was converted to a numeric type.
    - Several text columns (company_name, pay_period, job_posting_location, Work_type, Experience_level, skills_description, currency, compensation_type, company_industries_list, skills_list) were converted to lowercase and leading/trailing whitespace was removed. Missing values in these columns were filled with 'Not Specified'.
    - Rows with missing job_description were dropped.
    - The resume dataframe does not have any missing rows.
- **Text Cleaning:** Special characters and formatting issues were removed from job descriptions, skills descriptions, and resume text to create cleaned versions of these columns

final_yearly_salary still has a significant number of missing values (87,770 out of 123,849), which is expected as salary information is not always available for all job postings.

final_yearly_salary	87770
job_id	0
job_title	0
company_name	0
pay_period	0
job_posting_location	0
Work_type	0
job_description	0
remote_allowed	0
Experience_level	0
sponsored	0
skills_description	0
currency	0
compensation_type	0
company_industries_list	0
skills_list	0

Outlier Detection for Numerical Attributes

Outlier detection using the IQR method identified outliers in job_id (0.58% of non-NaN values) and final_yearly_salary (2.64% of non-NaN values). Outliers were removed from final_yearly_salary while keeping NaNs, resulting in a cleaned job_postings DataFrame used for subsequent analysis.

```
--- Outlier Analysis ---
```

	Attribute	Outlier Count	Outlier Percentage
0	job_id	718	0.58%
1	sponsored	0	0.00%
2	final_yearly_salary	952	2.64%

```
--- Outliers removed for final_yearly_salary ---
Original shape: (123842, 16)
Shape after removing outliers from 'final_yearly_salary': (122890, 16)
```

Statistical Summary for Numerical Attributes

Understanding the central tendency, dispersion, and shape of the distribution for numerical features is fundamental. The describe() function provides a concise summary of these characteristics.

Statistical analysis of numerical attributes in job_postings shows:

- job_id: A unique identifier.
- sponsored: Shows a standard deviation of 0, confirming all postings are not sponsored (value 0).
- final_yearly_salary: Shows significant variation. The average salary is approximately $205,328, with a median of $81,500. Salaries range from $0 to a maximum of $535,600,000. The 25th percentile is $52,000 and the 75th percentile is $125,000, indicating that most salaries fall below the mean, suggesting a right-skewed distribution influenced by high values.

```
--- Statistical Summary for Numerical Attributes ---
```

	job_id	sponsored	final_yearly_salary
count	1.238420e+05	123842.0	3.607200e+04
mean	3.896402e+09	0.0	2.053280e+05
std	8.404590e+07	0.0	5.097697e+06
min	9.217160e+05	0.0	0.000000e+00
25%	3.894587e+09	0.0	5.200000e+04
50%	3.901998e+09	0.0	8.150000e+04
75%	3.904707e+09	0.0	1.250000e+05
max	3.906267e+09	0.0	5.356000e+08

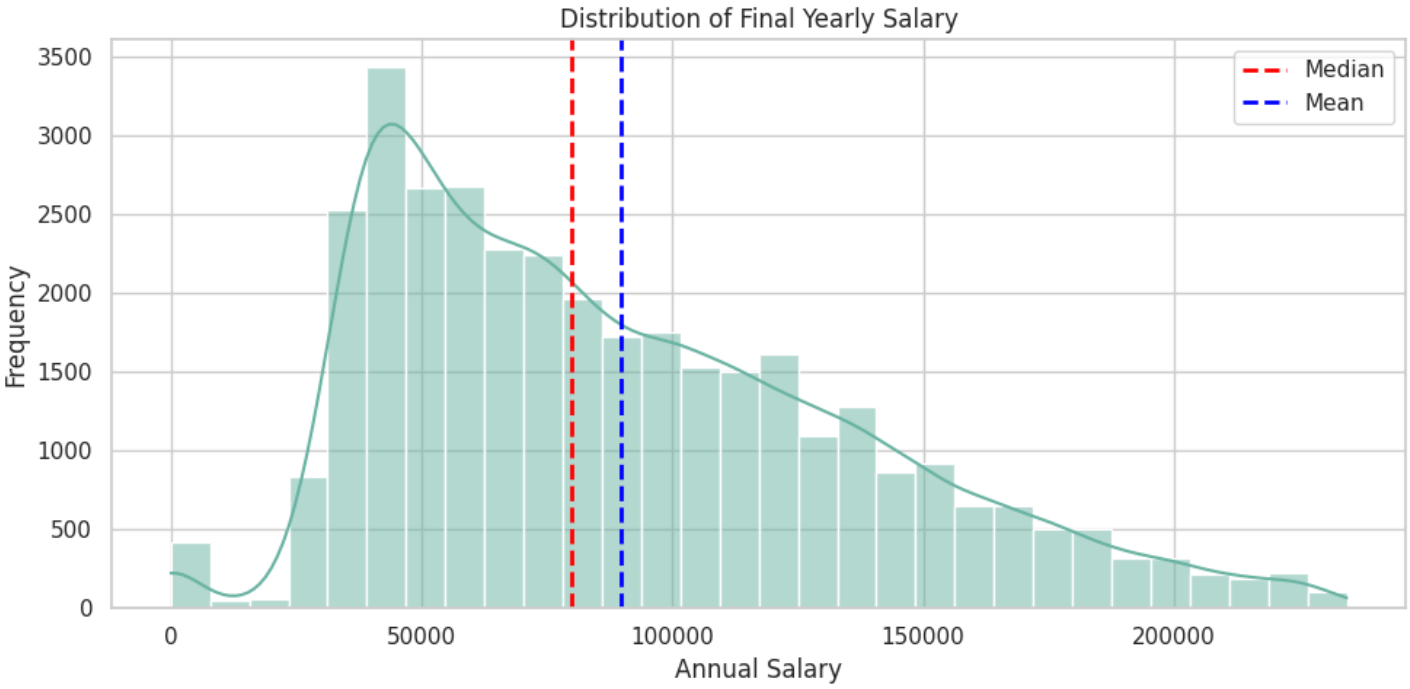
Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is crucial for uncovering patterns, anomalies, and relationships within the data. Python, with its rich ecosystem of data science libraries, was chosen for this purpose. The following packages were instrumental: Numpy for numerical operations, Pandas for data manipulation, Scikit-learn for potential future modeling utilities, Seaborn and Matplotlib for static visualizations, and Plotly for interactive plots.

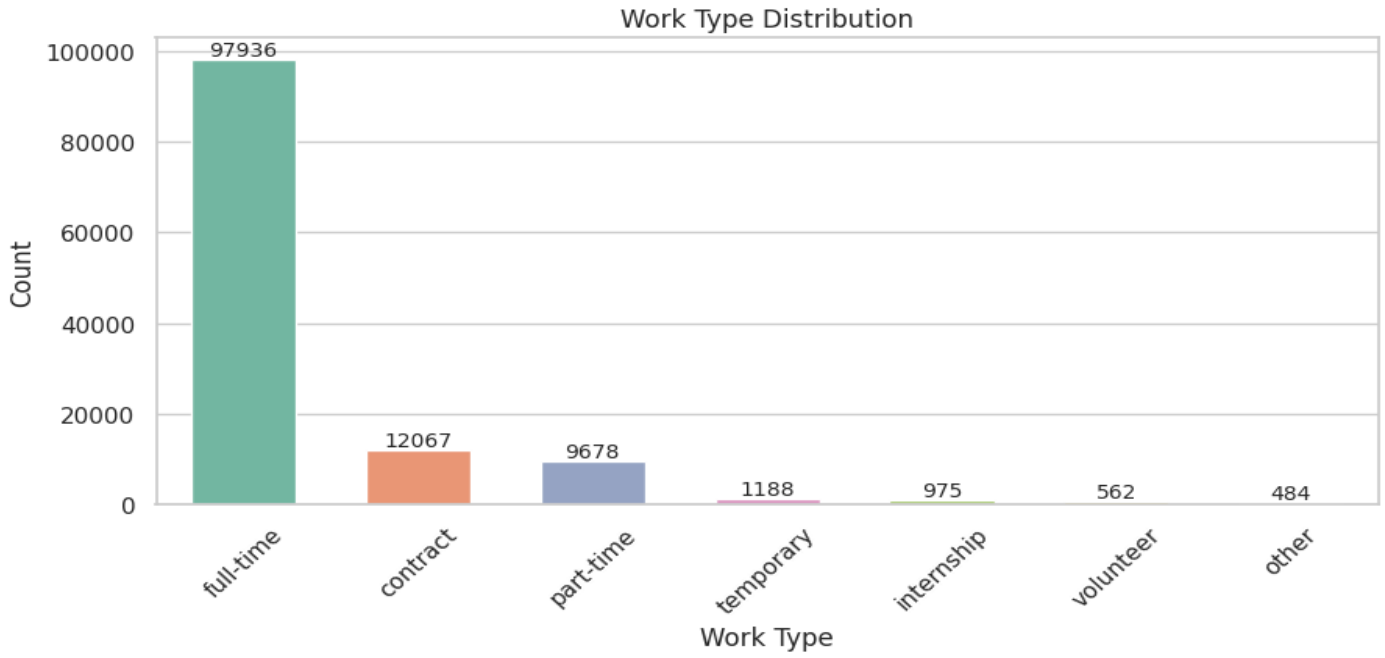
Univariate Analysis

Univariate analysis focuses on individual variables to understand their distributions, central tendencies, and spread. For our text-heavy dataset, this involved analyzing categories and numerical ranges.

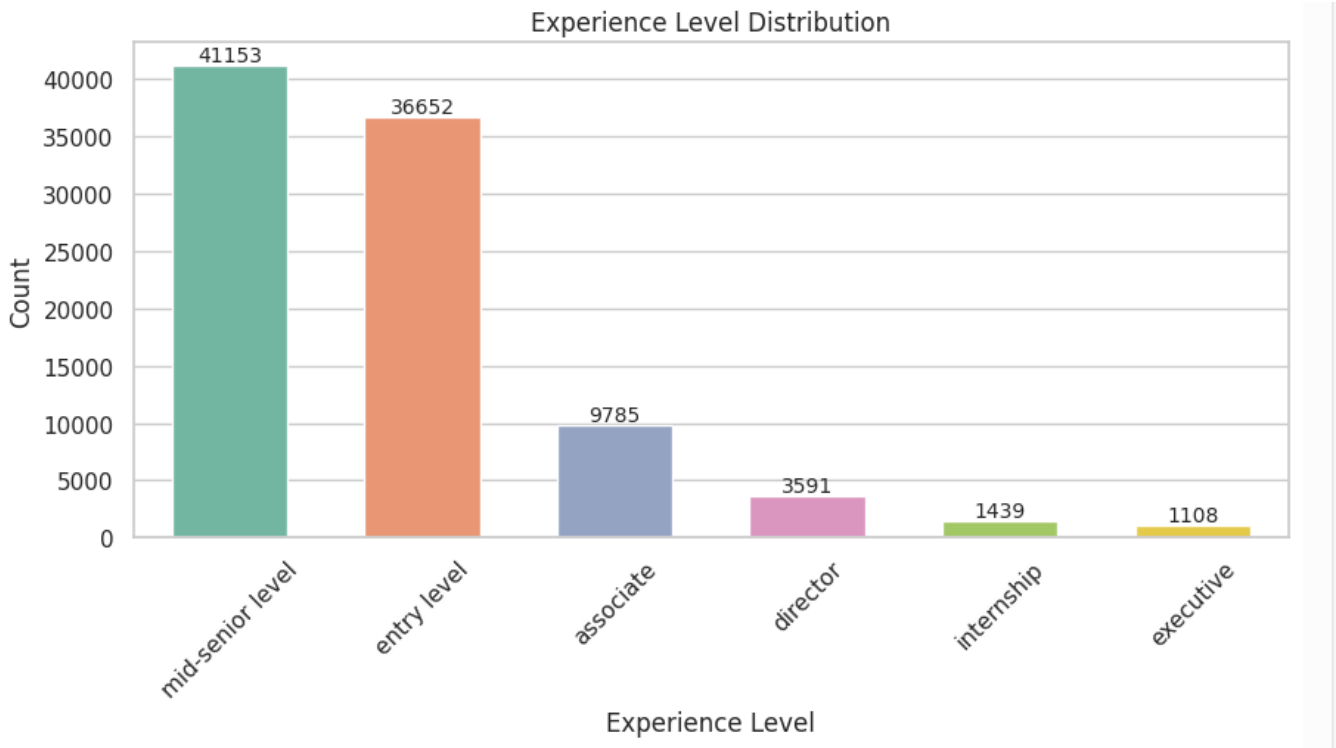
- **Distribution of Final Yearly Salary:** The histogram for final_yearly_salary clearly shows a right-skewed distribution. Most job postings cluster at lower salary ranges, with a long tail extending towards much higher salaries, visually confirming that the mean is pulled upwards by these higher values compared to the median.



- **Distribution of Work Type:** The count plot for Work type highlights the dominance of 'full-time' positions, with a significantly smaller number of 'contract' and 'part-time' roles. Other work types like temporary, internship, and volunteer are present in much lower frequencies.

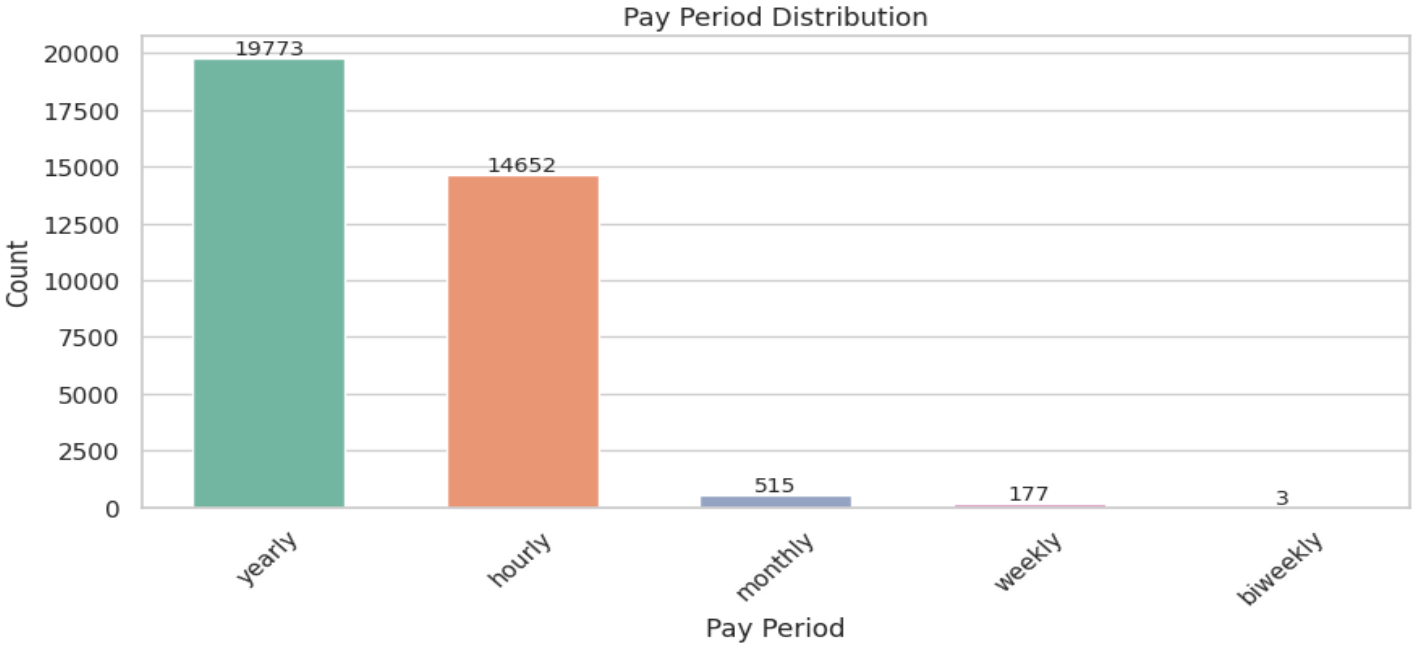


- **Distribution of Experience Level:** The distribution shows that 'mid-senior level' and 'Entry level' are the most common Experience Levels in the dataset. 'associate' is also frequent, while 'director', 'executive', and 'internship' are less common.

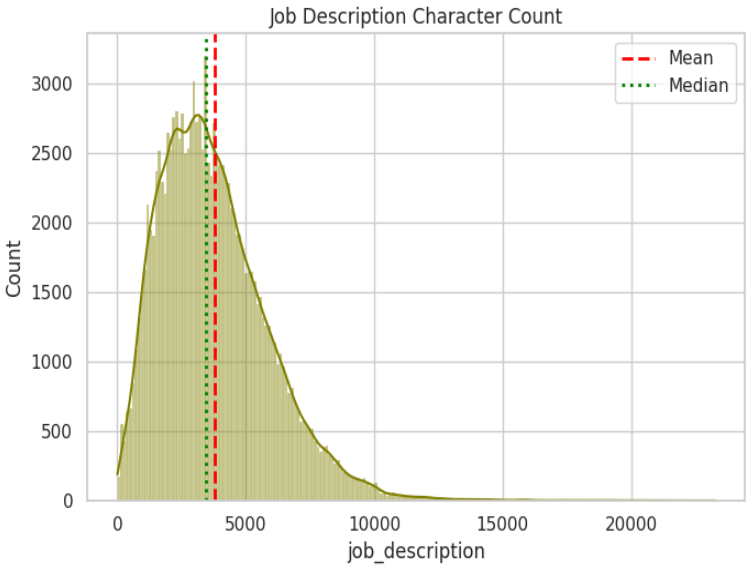


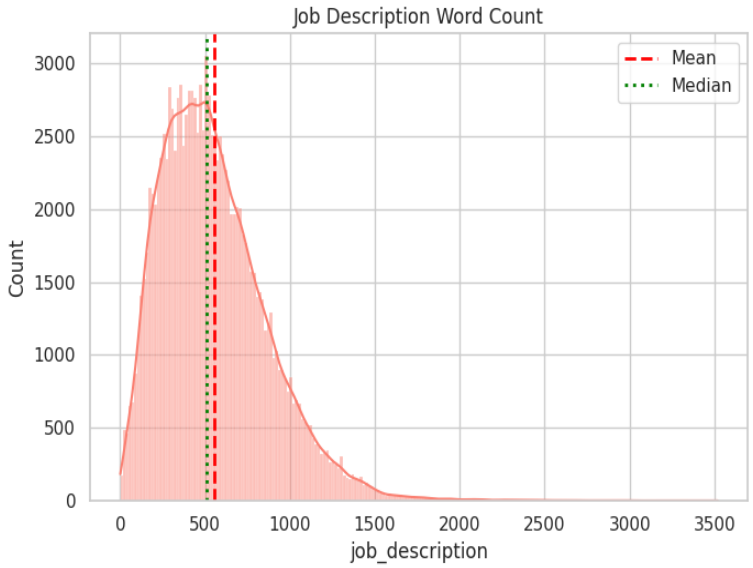
- **Distribution of Currency:** The plot clearly indicates that 'usd' is the primary currency used in the job postings, with other currencies being negligible.

- **Distribution of Pay Period:** The distribution shows that 'hourly' and 'yearly' are the most common pay periods, with 'monthly' being much less frequent.

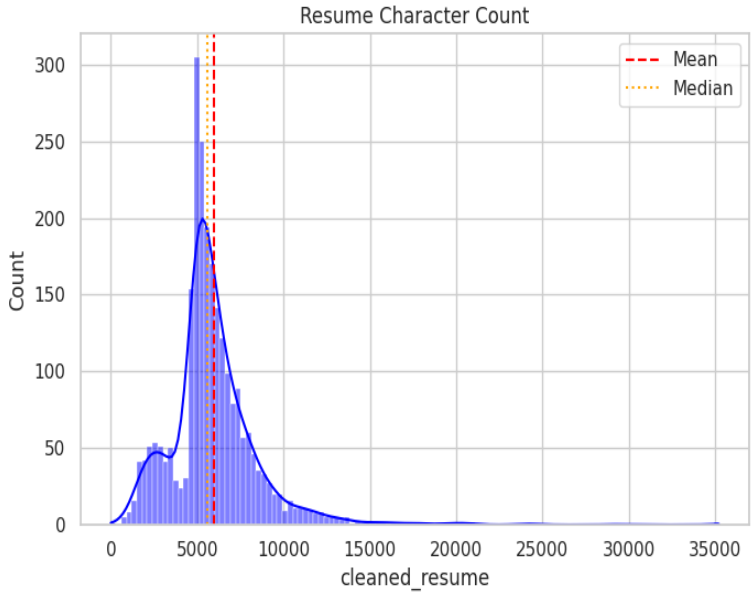


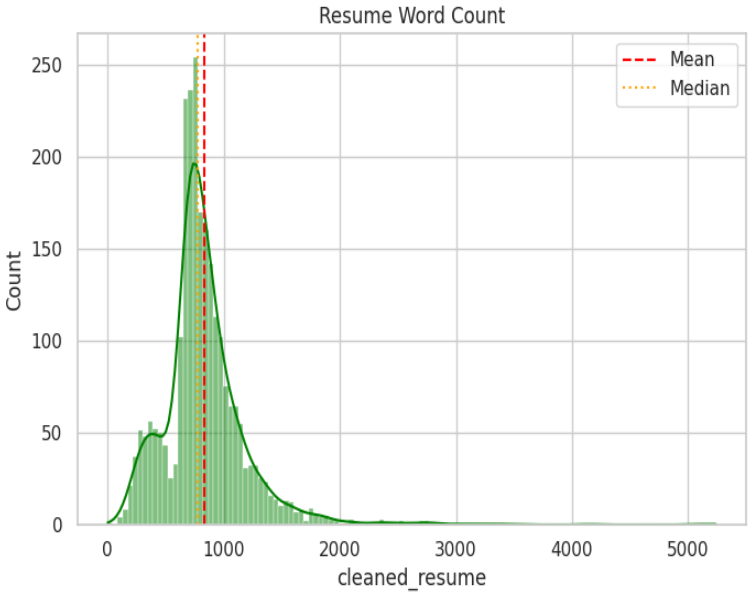
- **Job Description Character and Word Count Distribution:** The histograms for job description lengths (both character and word counts) show right-skewed distributions. This indicates that most job descriptions are relatively concise, with a smaller number of significantly longer descriptions, causing the mean length to be greater than the median length.



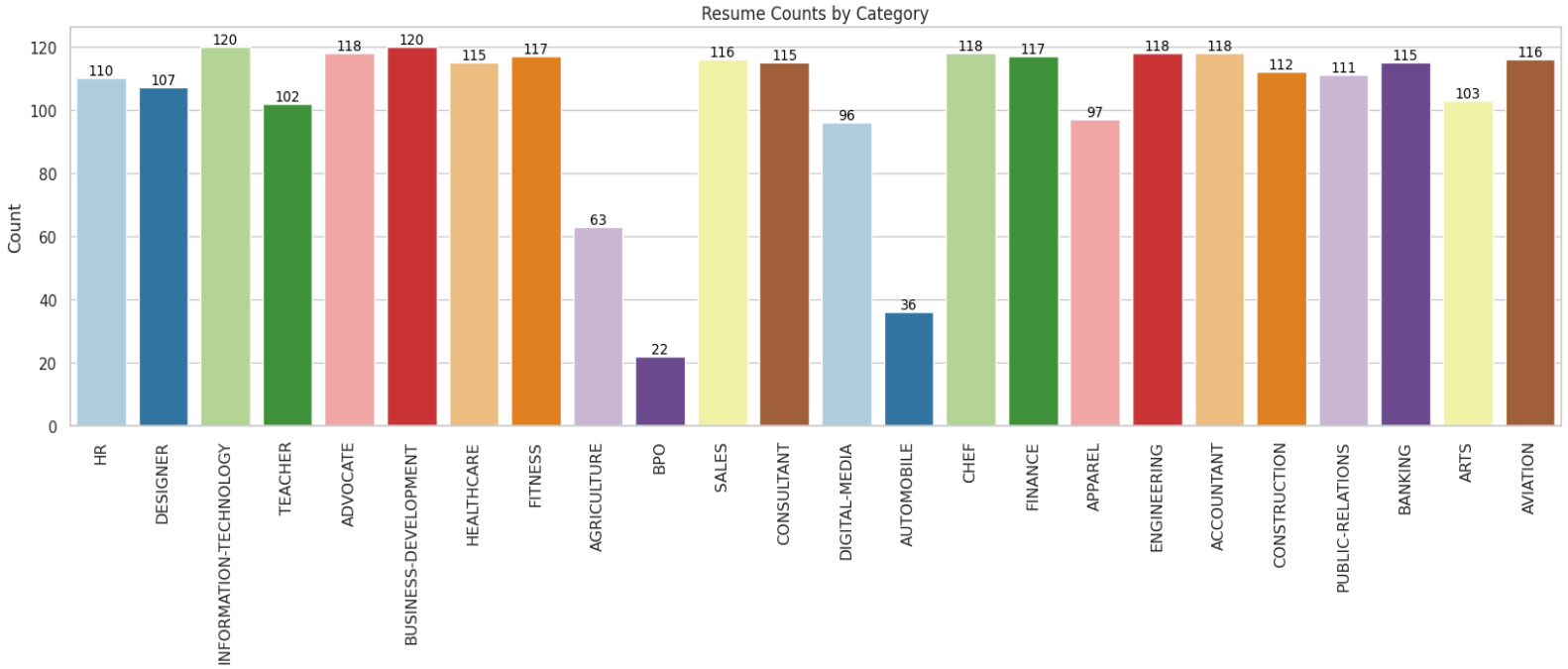


- **Resume Character and Word Count Distribution:** The histogram for resume word count exhibits a right-skewed distribution, suggesting that most resumes fall within a moderate length range, with fewer resumes being significantly longer.

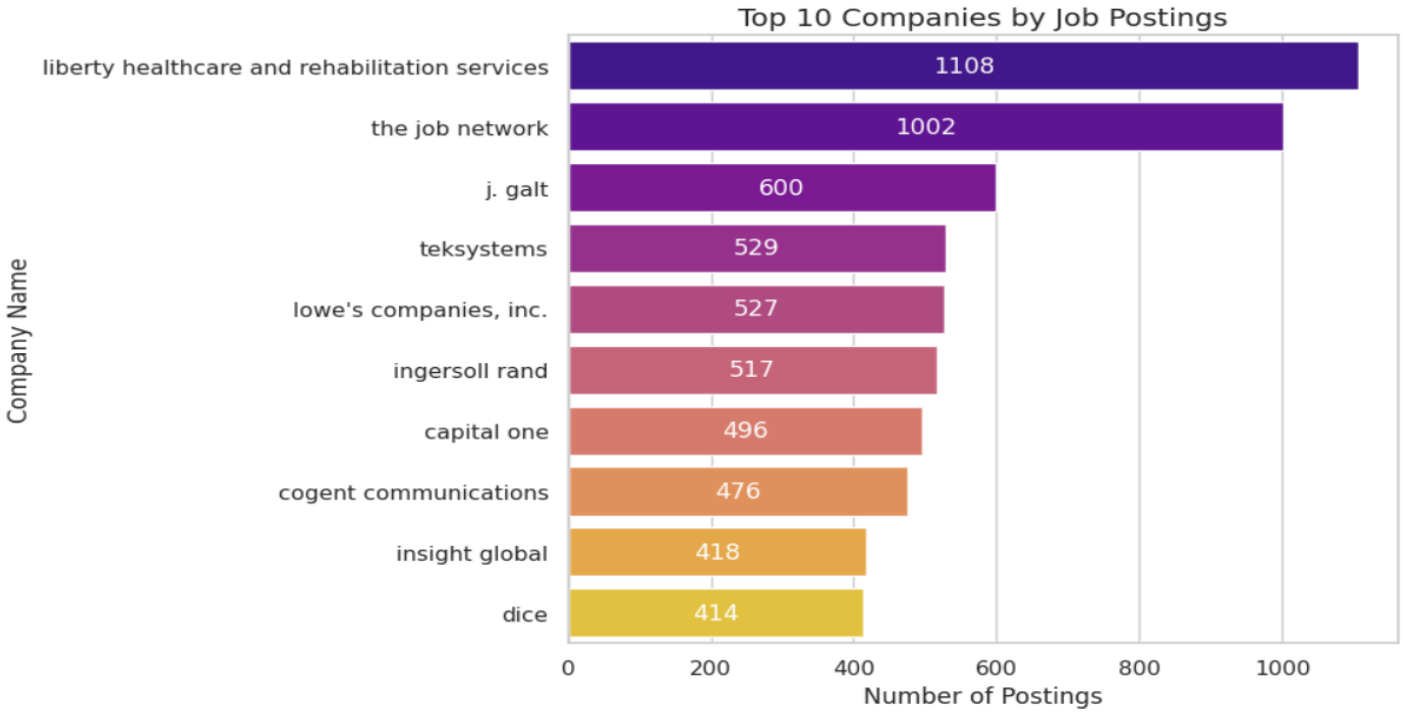




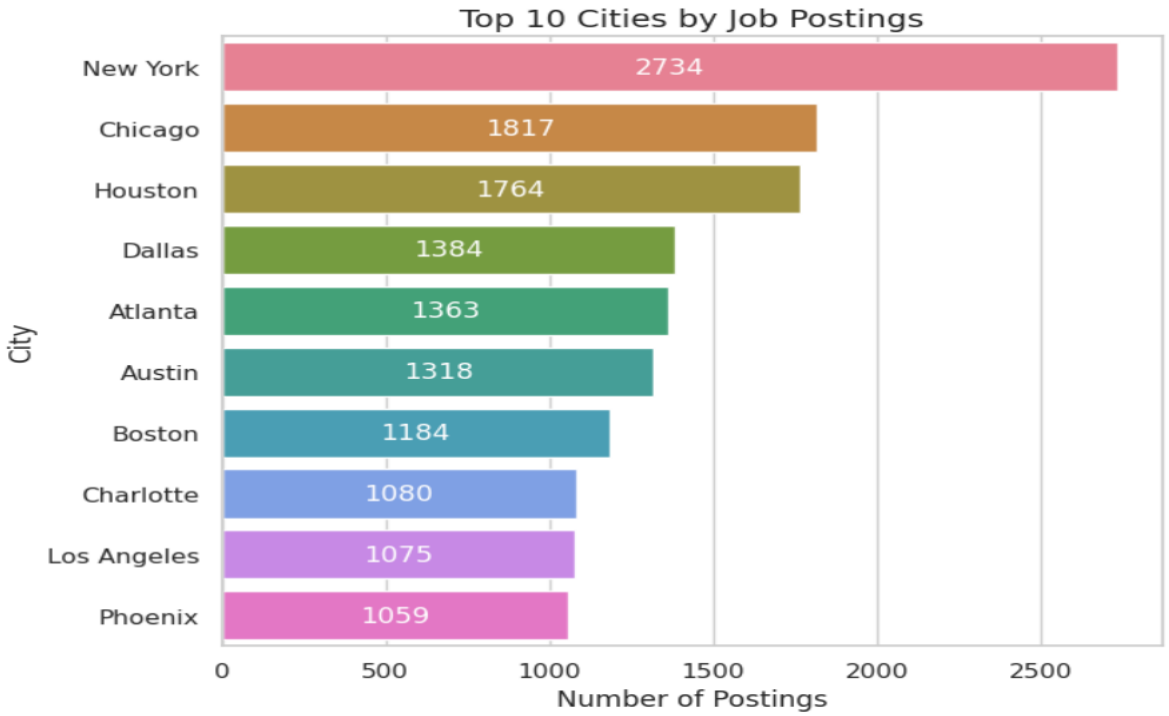
- **Resume distribution by Category:** The count plot for resume categories reveals the frequency of resumes across different professional fields. Categories like Information technology, Business development, Engineering and Accountant have the highest number of resumes in the resume dataset.



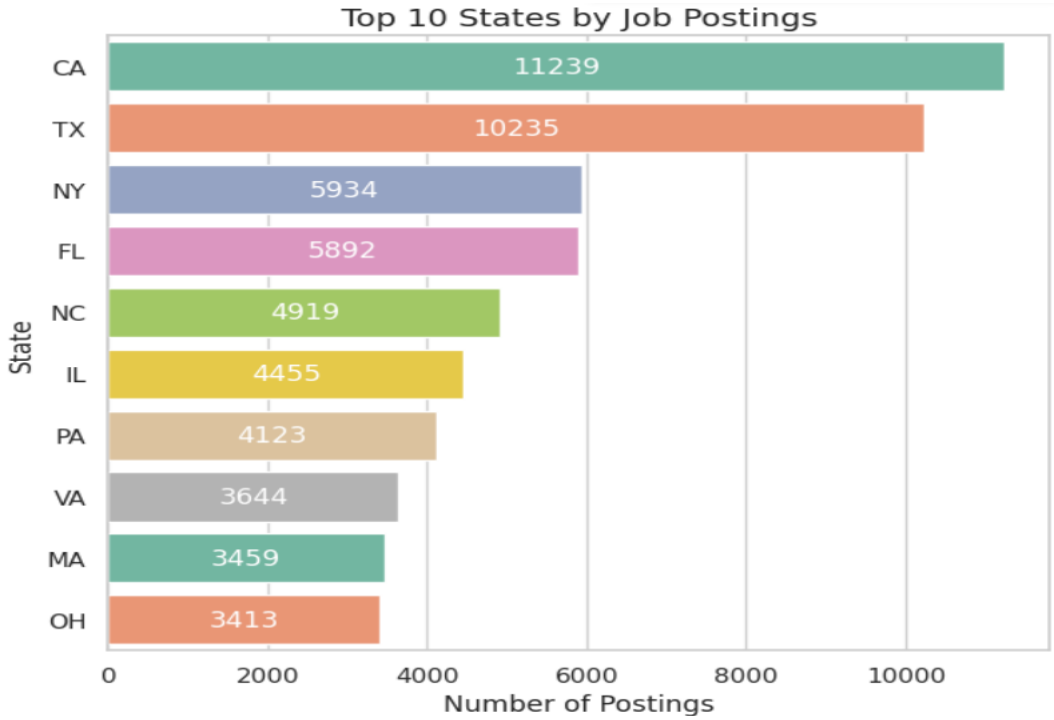
- **Top 10 Companies by Job Postings:** The count plot highlights the companies with the highest number of job postings.



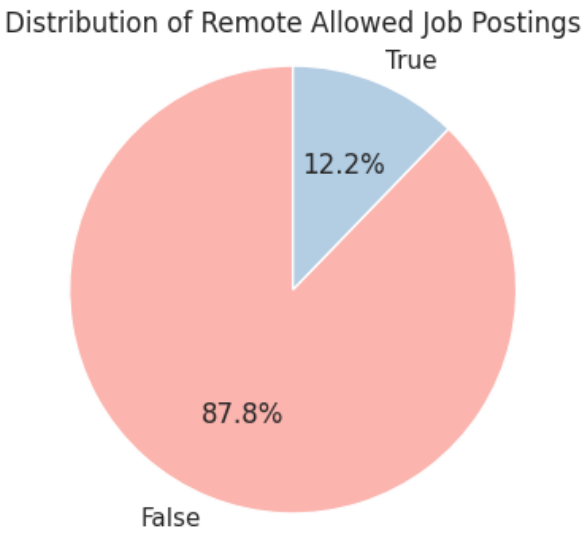
- **Top 10 Cities by Job Postings:** The count plot identifies cities with the highest number of job postings. Cities like New York, Chicago, Houston shows highest number of job postings.



- **Top 10 States by Job Postings:** The count plot identifies states with the highest job posting counts, which are California, Texas and New York.



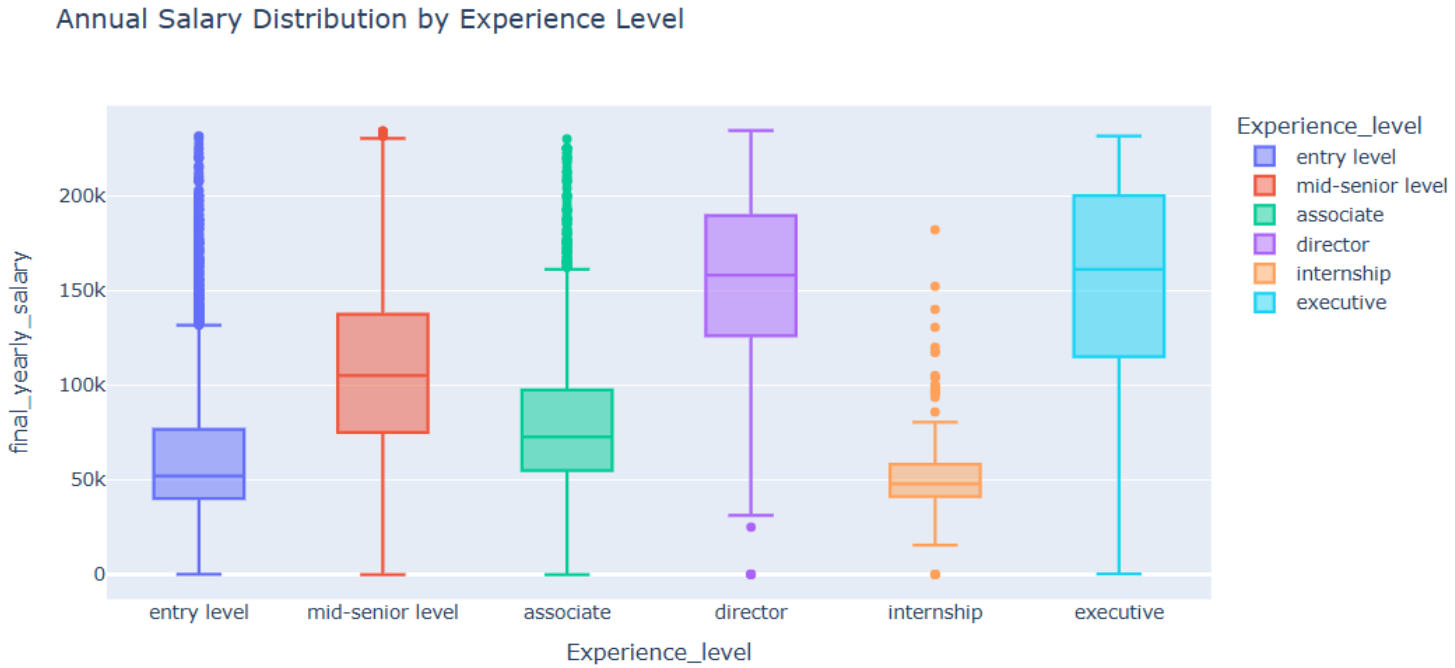
- **Distribution of Remote Allowed Job Postings:** The pie plot shows the proportion of jobs that are remote vs. not remote. A significant majority (around 87.8%) are not remote, while around 12.2% allow remote work.



Bivariate Analysis

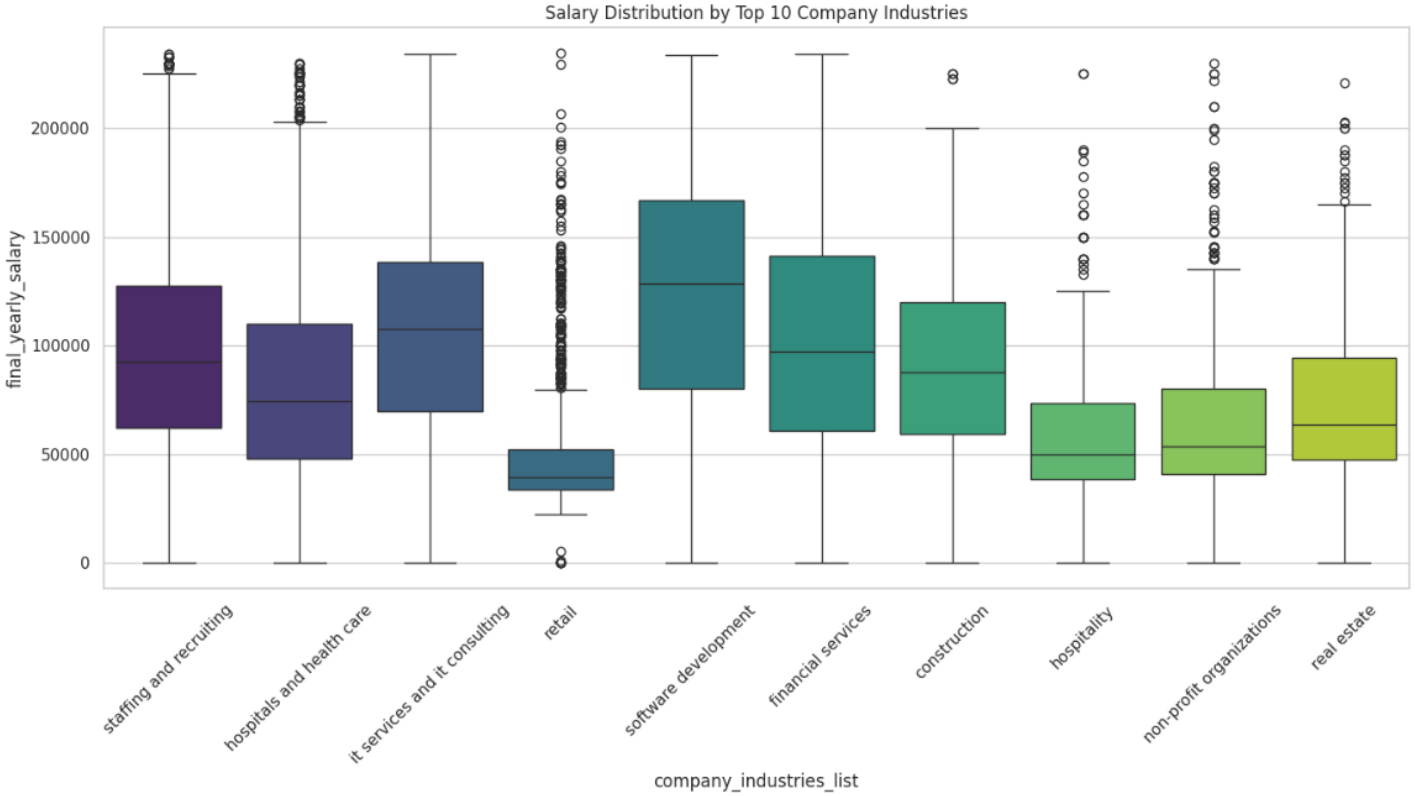
Bivariate analysis explores the relationships between two variables. This helps in understanding how different job posting attributes relate to each other, which can inform our feature selection and similarity modeling.

- **Salary Distribution by Experience Level:** The box plot shows a clear trend where median salaries generally increase with Experience Level. 'Executive' and 'director' levels have the highest median salaries and wider salary ranges, while 'entry level' and 'internship' have lower medians and narrower ranges.



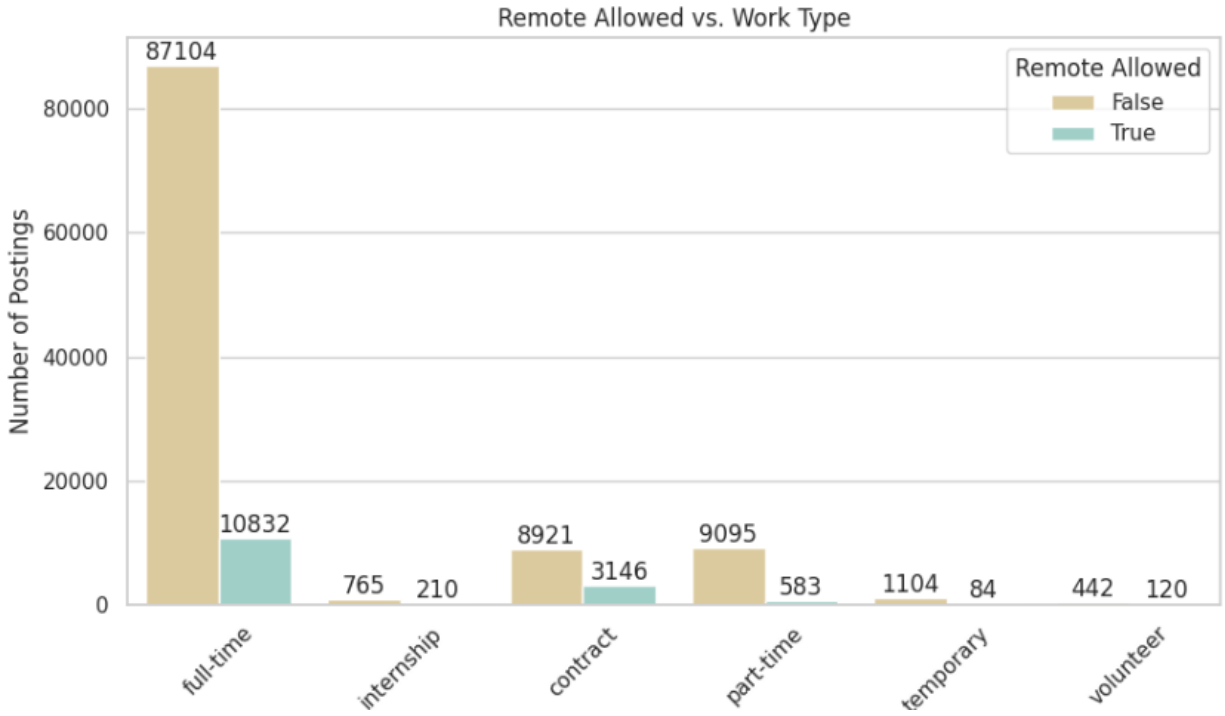
- **Salary Distribution by Top 10 Industries:** The box plot reveals variations in salary distributions across different top industries. Some industries, like 'software development' and 'It

services and It consulting', tend to have higher median salaries compared to others like 'retail' and 'hospitality'.

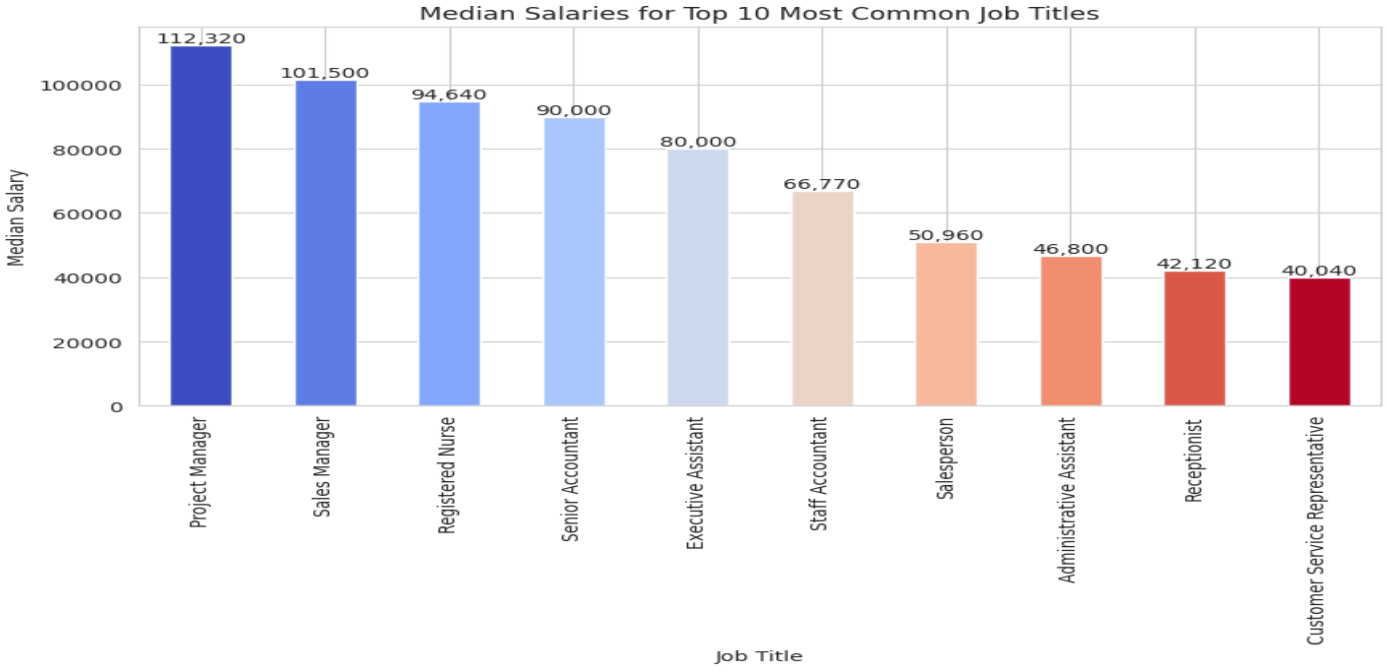


- **Remote Allowed vs. Experience Level:** The count plot with hue shows that while remote jobs exist across experience levels, the proportion of remote jobs varies. 'Mid-senior level' and 'entry level' have a notable number of remote postings, but 'executive' and 'internship' have fewer remote options.

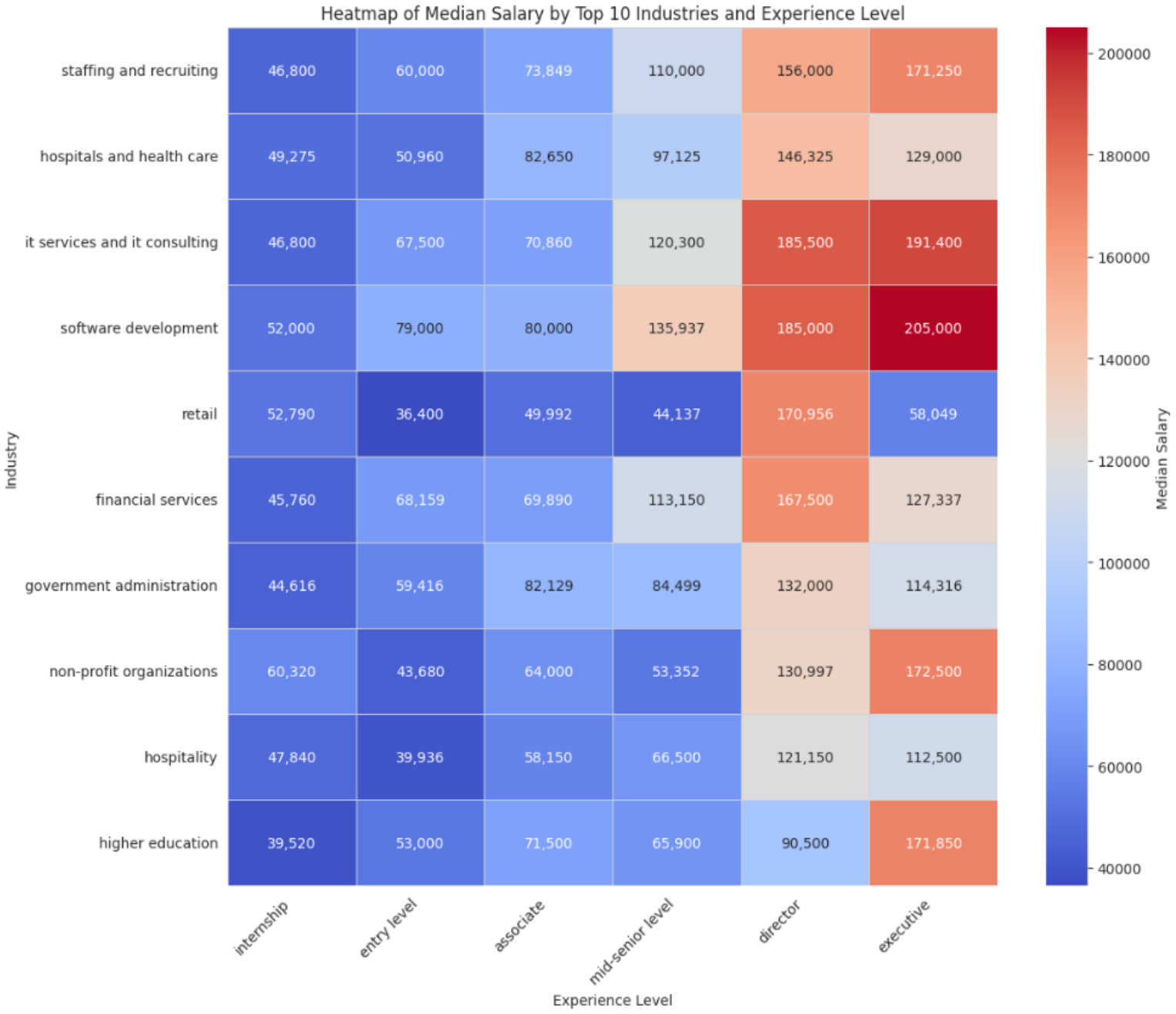
- **Remote Allowed vs. Work Type:** The count plot with hue shows that remote work is most prevalent in 'full-time' positions, although remote options also exist for 'contract' and 'part-time' roles, albeit in much smaller numbers.



- **Median Salaries for Top 10 Most Common Job Titles:** The bar plot explicitly shows the median salary for the ten most frequent job titles. Job titles like 'Project Manager' and 'Sales Manager' show higher median salaries compared to titles like 'Administrative Assistant' or 'Receptionist'.



- **Heatmap of Median Salary by Top 10 Industries and Experience Level:** This heatmap provides a detailed view of median salary variations. It shows that within the top industries; median salaries generally increase with experience level. High-paying combinations include industries like 'software development' and 'IT services and IT consulting' at 'mid-senior level', 'director', and 'executive' levels, with specific median salary figures visible on the heatmap cells. Industries like 'retail' or 'non-profit organizations' tend to show slightly lower median salaries across experience levels.



References

1. Bevara, R. V. K., Mannuru, N. R., Karedla, S. P., Lund, B., Xiao, T., Pasem, H., Dronavalli, S. C., & Rupeshkumar, S. (2025). *Resume2Vec: Transforming Applicant Tracking Systems with Intelligent Resume Embeddings for Precise Candidate Matching*. Electronics (Basel), 14(4), 794–. https://doi.org/10.3390/electronics14040794

2. Ayishathahira, C. H., Sreejith, C., & Raseek, C. (2018). *Combination of Neural Networks and Conditional Random Fields for Efficient Resume Parsing*. 2018 International CET Conference on Control, Communication, and Computing (IC4), 388–393. https://doi.org/10.1109/CETIC4.2018.8530883

3. Bhatt, A., Mittal, S., Uniyal, A., Tiwari, P., Jyala, D., & Singh, D. (2024). *Resume Analyzer based on MapReduce and Machine Learning*. 2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI), 2, 1–5. https://doi.org/10.1109/IATMSI60426.2024.10503467

4. Burrell, D. N., & Mcandrew, I. (2023). *Exploring the Ethical Dynamics of the Use of Artificial Intelligence (AI) in Hiring in Healthcare Organizations*. Land Forces Academy Review, 28(4), 309–321. https://doi.org/10.2478/raft-2023-0037

5. Delecraz, S., Eltarr, L., Becuwe, M., Bouxin, H., Boutin, N., & Oullier, O. (2022). *Making recruitment more inclusive: unfairness monitoring with a job matching machine-learning algorithm*. 2022 IEEE/ACM International Workshop on Equitable Data & Technology (FairWare), 34–41. https://doi.org/10.1145/3524491.3527309

6. Fettach, Y., Bahaj, A., & Ghogho, M. (2025). *Skill Demand Forecasting Using Temporal Knowledge Graph Embeddings*. https://doi.org/10.48550/arxiv.2504.07233

7. Frissen, R., Adebayo, K. J., & Nanda, R. (2023). *A machine learning approach to recognize bias and discrimination in job advertisements*. AI & Society, 38(2), 1025–1038. https://doi.org/10.1007/s00146-022-01574-0

8. Gaur, B., Saluja, G. S., Sivakumar, H. B., & Singh, S. (2021). *Semi-supervised deep learning based named entity recognition model to parse education section of resumes*. Neural Computing & Applications, 33(11), 5705–5718. https://doi.org/10.1007/s00521-020-05351-2

9. J, P. V., P, S. N. J., Gopinath, S., S, U., & C.R., K. (2024). *Resume Analyzer and Skill Enhancement Recommender System*. 2024 Asia Pacific Conference on Innovation in Technology (APCIT), 1–6. https://doi.org/10.1109/APCIT62007.2024.10673530

10. Kinger, S., Kinger, D., Thakkar, S., & Bhake, D. (2024). *Towards smarter hiring: resume parsing and ranking with YOLOv5 and DistilBERT*. Multimedia Tools and Applications, 83(35), 82069–82087. https://doi.org/10.1109/s11042-024-18778-9

11. Özcan, G., & Ögüdücü, S. G. (2017). *Applying Classifications Techniques in Job Recommendation System for Matching of Candidates and Advertisements*. International Journal

of Intelligent Computing Research, 8(1), 798–806. https://doi.org/10.20533/ijicr.2042.4655.2017.0098

12. Pabalkar, S., Patel, P., Choudhary, R., Panoch, V., Yadav, S., & Ghogale, H. (2024). *Resume Analyzer Using Natural Language Processing (NLP)*. 2024 International Conference on Intelligent Systems and Advanced Applications (ICISAA), 1–6. https://doi.org/10.1109/ICISAA62385.2024.10828940
13. Singh, S., & Gupta, P. (2023). *Talent Recommendation Engine with Real-Time Feedback Loop*. Proceedings of the Third International Conference on AI-ML Systems, 1–4. https://doi.org/10.1145/3639856.3639911
14. Sruthi, P., Adithya, P. N. V. K. G., Suleman, M. D., Kunal, P., & Gairola, S. P. (2023). *Smart Resume Analyser: A Case Study using RNN-based Keyword Extraction*. E3S Web of Conferences, 430, 1023–. https://doi.org/10.1051/e3sconf/202343001023
15. Zhang, L., & Yencha, C. (2022). *Examining perceptions towards hiring algorithms*. Technology in Society, 68, 101848–. https://doi.org/10.1016/j.techsoc.2021.101848
16. Zhao, J., Wang, J., Sigdel, M., Zhang, B., Hoang, P., Liu, M., & Korayem, M. (2021). *Embedding-based Recommender System for Job to Candidate Matching on Scale*. https://doi.org/10.48550/arxiv.2107.00221

Dataset

The project uses two publicly available datasets:

1. **Resume Dataset** (Kaggle):
    - Contains 2485 resumes.
    - Source: Resume Dataset.
2. **Job Postings Dataset** (Kaggle):
    - Includes 123,849 records of job postings
    - Source: Job Postings Dataset.

GitHub Repository link

https://github.com/rabiadanish/ResumeMatch