IND5003 Data Analytics for Sense-making

LinkedIn Data Analytics Report

Project Group 3

Wang Mingyang e1582388@u.nus.edu

2025-10-26

Abstract

This study examines LinkedIn job postings from 2023–2024 to uncover patterns in job demand, salary transparency, and the economic value of skills. Following comprehensive data cleaning and integration, a combination of descriptive statistics and machine learning methods was employed to analyze how job roles, skills, and benefits influence labor market behavior. The analysis was structured around four research questions: identifying dominant job and skill trends (RQ1), evaluating salary disclosure across industries and work types (RQ2), discovering latent occupational structures through unsupervised clustering (RQ3), and investigating how individual and combined skills relate to compensation (RQ4). Text mining techniques, including TF-IDF vectorization, Truncated SVD, PCA, and K-Means clustering, were used to extract semantic structures from unstructured job text data, while correlation and uplift analyses quantified pay differentials. The findings reveal that interdisciplinary skill combinations—such as Engineering + Finance or Marketing + Sales—yield notable salary premiums, whereas overall skill popularity exhibits only weak association with compensation. These results emphasize the rising market value of cross-domain expertise and demonstrate how data-driven text analytics can provide actionable insights into the structure and dynamics of the digital labor economy.

Keywords: LinkedIn job postings; salary transparency; skill analytics; unsupervised clustering; TF-IDF; K-Means; salary premium; labor market intelligence.

1. Introduction

The transformation of the global labor market through digital platforms has enabled the systematic analysis of large-scale employment data. Among these platforms, LinkedIn stands out as one of the most extensive sources of professional and recruitment information, encompassing millions of job postings, company profiles, and user interactions. Such data have become instrumental in understanding the evolving relationships between job roles, required skills, and compensation in an increasingly knowledge-driven economy. The emergence of data-driven recruitment has not only improved hiring efficiency but also opened new opportunities for empirical research in labor market analytics. By analyzing large datasets of job postings, researchers can uncover structural trends in occupational demand, salary transparency, and skill evolution across industries.

In this context, LinkedIn job postings data serve as a rich empirical foundation for studying how digital labor markets reflect economic activity and organizational needs. The dataset used in this study—LinkedIn Job Postings (2023–2024) compiled by Arsh Koneru on Kaggle[1], which contains hundreds of thousands of postings collected from multiple global regions, with attributes covering job titles, company information, benefits, salaries, skills, and industry affiliations.

The purpose of this research is to perform a comprehensive, data-driven exploration of the LinkedIn employment landscape during 2023–2024. This study aims to identify major market signals across occupations, analyze salary transparency trends, uncover hidden job categories using unsupervised learning, and quantify the relationship between skill composition and compensation. Through this analysis, the study seeks to reveal both observable and latent dimensions of the modern job market, providing insights for data scientists, recruiters, and policymakers alike.

The research is guided by four central objectives. The first is to analyze market signals across job roles, industries, and skills to identify dominant employment patterns and benefit structures. The second is to investigate the degree of salary transparency and how it differs across sectors and job types. The third explores whether unsupervised learning models, such as TF-IDF, TruncatedSVD, and KMeans clustering, can uncover hidden job categories based on textual and semantic similarities. The fourth examines how skill demand translates into compensation, with a particular focus on whether combinations of skills produce measurable salary uplift compared to single-skill baselines.

Methodologically, the project integrates exploratory data analysis (EDA) with natural language processing (NLP) and machine learning. This hybrid analytical framework allows for both descriptive insight and model-driven inference, bridging statistical reasoning with computational approaches to textual data. By combining frequency-based skill analysis, clustering, and correlation testing, this study demonstrates how data science techniques can contribute to computational labor economics, offering a quantitative perspective on workforce trends.

In essence, this report contributes to understanding the dynamics of digital employment ecosystems. It highlights how interdisciplinary skill sets—especially those combining technical and business expertise—are increasingly rewarded in the modern

economy. The findings ultimately emphasize that large-scale online job data are not merely administrative records but valuable signals of broader structural changes in how work, skills, and value are produced in a connected world.

2. Method Overview

This study employs an integrated analytical framework that combines data preprocessing, exploratory visualization, natural language processing (NLP), and unsupervised learning to extract meaningful labor market insights from LinkedIn job postings. The methodology emphasizes both reproducibility and interpretability, reflecting best practices in computational social science research. The analysis was implemented in Python using core data science libraries such as pandas, scikit-learn, matplotlib, and seaborn within a Jupyter Notebook environment.

The first stage of the analysis involved data cleaning and normalization. Multiple tables describing jobs, companies, industries, skills, and benefits were merged through unique identifiers, and text fields were preprocessed using regular expressions to remove redundancy and standardize format. Non-informative entries such as "Not Specified" were filtered out to ensure analytical consistency. The resulting dataset provided a coherent structure suitable for downstream statistical and text-based modeling.

In the exploratory phase, descriptive statistics and visualizations were generated to identify the most frequent job titles, industries, and skill terms. This step, guided by exploratory data analysis (EDA) principles, allowed for a preliminary understanding of the dataset's internal composition and informed the design of subsequent research questions. Techniques such as bar plots, histograms, and word clouds were applied to visualize dominant employment patterns and benefit distributions.

The next stage focused on text representation and semantic pattern discovery. Job descriptions were converted into numerical vectors using the Term Frequency–Inverse Document Frequency (TF-IDF) method[2], which quantifies the importance of words relative to the entire corpus. To address the high dimensionality inherent in TF-IDF matrices, Truncated Singular Value Decomposition (TruncatedSVD) [3] was applied as a linear dimensionality reduction technique, producing dense, low-rank embeddings suitable for clustering and visualization. The reduced embeddings were then grouped using KMeans clustering[4], which partitions samples into k clusters by minimizing within-cluster variance, effectively identifying latent thematic structures in the job data. Six major clusters were observed, corresponding to occupational domains such as engineering, finance, healthcare, and sales. To facilitate visual interpretation, Principal Component Analysis (PCA) was further used to project these embeddings into a two-dimensional space, revealing the separability of job categories while maintaining global variance structure.

The final analytical component examined salary-related relationships. Salary fields were normalized into annualized values, allowing for cross-posting comparison. To assess whether skill popularity correlates with compensation, Pearson's correlation coefficient [5] was computed, quantifying the strength and direction of linear association between skill frequency and mean salary. Furthermore, a skill-pair uplift analysis was introduced to

measure how certain combinations of skills influence salary outcomes relative to individual baselines. This approach represents an extension of standard correlation analysis by capturing potential synergistic effects across co-occurring skill sets. The implementation of this analysis, while not covered in the course material, reflects an element of self-directed learning and methodological innovation, demonstrating how statistical logic can be extended to novel data contexts.

Visualization played a central role throughout the study. Seaborn and Matplotlib were employed for static analytical plots, while Plotly enabled interactive visual exploration of salary distributions, benefit patterns, and skill networks. These visualization strategies align with modern standards of transparent and reproducible analysis, where graphical summaries serve not only as illustrative tools but as a means of iterative hypothesis refinement. The code structure followed a modular design, mapping each analysis segment directly to a corresponding research question. This workflow ensures traceability, clarity, and extensibility for high-quality data analysis.

3. Research Questions of Interest

The analysis in this report is guided by four research questions, each addressing a specific dimension of the LinkedIn job postings dataset. Together, these questions aim to build a holistic understanding of how job characteristics, skills, and compensation interact within the digital labor market. Each question is grounded in observable data patterns and implemented through the analytical framework described in the previous chapter.

RQ1 focuses on identifying market signals embedded in job roles, skills, and benefits. The purpose of this inquiry is to understand which job titles dominate the dataset, which skills appear most frequently, and how employers use benefits to enhance job attractiveness. By examining frequency distributions, word clouds, and aggregated indicators, this stage provides a foundation for interpreting broader employment trends. It also establishes the contextual basis for more advanced analyses in subsequent sections.

RQ2 investigates salary transparency and its variation across industries and work types. This part of the study explores the extent to which employers disclose salary information and how compensation levels differ between technical, healthcare, and service-oriented sectors. Such analysis provides insight into both organizational behavior and the evolving norms of pay disclosure. By quantifying transparency and analyzing salary ranges, this stage contributes to understanding the structural and ethical dimensions of digital recruitment.

RQ3 aims to uncover hidden thematic clusters among job postings using unsupervised learning. Job descriptions are transformed into numerical representations through TF-IDF vectorization and dimensionality reduction, followed by KMeans clustering. This approach captures latent structures that are not apparent in descriptive statistics, revealing distinct occupational domains such as engineering, finance, healthcare, and sales. By visualizing these clusters in a reduced semantic space, the analysis enhances interpretability and offers an empirical basis for classifying jobs according to their textual and conceptual similarities.

RQ4 examines the relationship between skills and compensation, with a particular emphasis on whether certain skill combinations yield higher salary uplift compared to

individual skills. This analysis quantifies how skill composition influences earnings and tests whether interdisciplinary capabilities—such as combining business knowledge with technical expertise—provide measurable advantages in the labor market. By comparing skill-pair outcomes with baseline averages, the results provide empirical evidence of the economic value of cross-domain expertise.

Overall, these four research questions are designed to progress from descriptive understanding to structural discovery and quantitative validation. The sequence reflects an iterative analytical philosophy: each question refines the insights of the previous one while expanding the scope of inquiry. Collectively, they enable a comprehensive exploration of job market behavior grounded in both data-driven methods and interpretive reasoning.

4. Data Overview and Preparation

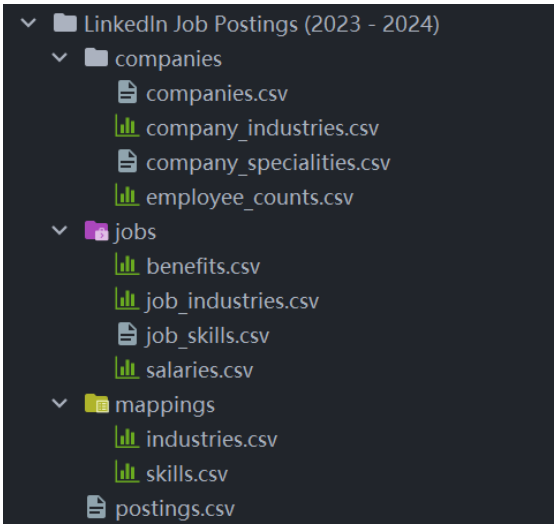

Figure.1. Dataset folder structure

The LinkedIn Job Postings dataset used in this study is organized into three main folders: companies, jobs, and mappings. The companies folder contains corporate-level metadata, including company descriptions, industry affiliations, specialties, and workforce size. The jobs folder contains job-level information such as benefits, industries, skills, and salaries. Finally, the mappings folder provides lookup tables that link unique industry and skill identifiers to their textual descriptions. A single master file named postings.csv consolidates all postings and serves as the central table for analysis.

```
cols_order = [
    # keys
    "job_id", "company_id",
    # job
    "title", "description", "location", "formatted_work_type", "remote_flag_simple",
    "formatted_experience_level", "skill_names", "industry_names", "benefits",
    # salary
    "pay_period", "currency", "compensation_type",
    "min_salary", "med_salary", "max_salary", "avg_salary",
    # company
    "name", "company_size", "country", "city", "address", "url",
    "employee_count", "follower_count",
    "company_industries_list", "company_specialities_list",
    # misc
    "posting_domain", "application_url", "applies", "views",
    "remote_allowed", "time_recorded"
]
```

Figure.2. Some key dataset fields

Each posting record corresponds to an individual job advertisement uniquely identified by a job ID and linked to its corresponding company ID. Supporting attributes are grouped into four logical categories: job-specific details (such as title, description, work type, and experience level), salary information (minimum, median, and maximum salary values), company characteristics (including company size, location, and number of employees), and miscellaneous attributes (such as posting domain, application link, number of views, and timestamp).This schema allows a relational join across the supporting tables to construct a unified analytical dataset suitable for descriptive, textual, and predictive analyses.

During the integration stage, tables from the companies and jobs folders were merged based on their shared identifiers. Records containing incomplete or missing identifiers were excluded to prevent duplication or mismatch. Each column was converted into an appropriate data type (numerical for salary-related fields), categorical for work types and experience levels, and text for skills and benefits. The final merged dataset consisted of hundreds of thousands of valid postings containing consistent and harmonized attributes ready for further analysis.

Preprocessing focused on ensuring semantic consistency across fields. For textual attributes such as industry name and benefits, standardization was necessary due to inconsistent punctuation, duplicated entries, and mixed capitalization. Regular expression transformations were applied to correct spacing issues, unify synonyms, and remove ambiguous conjunctions (for instance, merging “Oil, Gas, and Mining” with “Oil and Gas”). For skill entries, the text strings were split into lists of distinct skills after trimming whitespace and removing non-informative values such as “Not Specified.” Each job record was then reduced to a unique skill set, ensuring accurate aggregation and later enabling analyses such as skill frequency counts and co-occurrence pair construction.

The cleaned dataset therefore serves as a well-structured foundation for the analytical workflow described in subsequent chapters. It integrates information across multiple relational tables while maintaining interpretability and reproducibility.

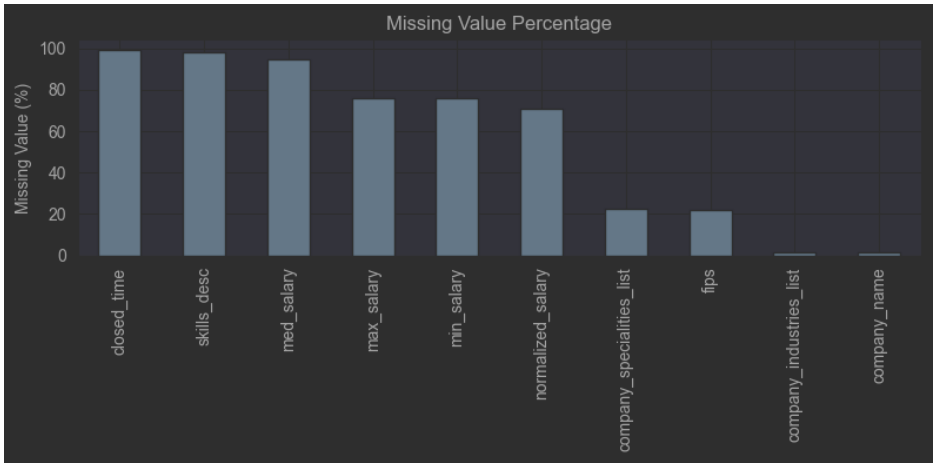

Figure.3. Missing Value Pecentage

To validate data integrity after preprocessing, a brief exploratory visualization was conducted. Figure 3 presents the proportion of missing values across major attributes. While certain auxiliary columns, such as closed_time and skills_desc, exhibit high rates of missingness due to incomplete LinkedIn postings, the essential analytical variables—including job identifiers, titles, work types, and company information—are well populated, confirming the reliability of the core dataset.

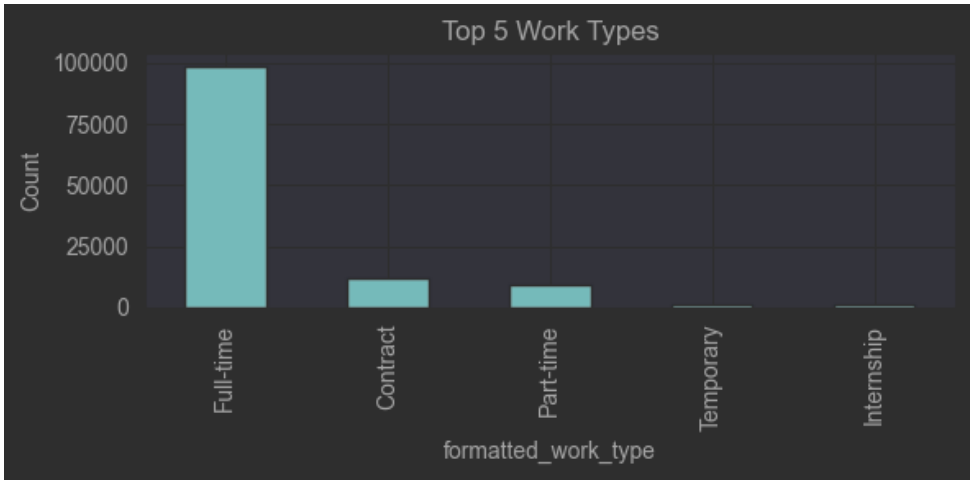

Figure.4. Top 5 Work Types

Figure 4 further illustrates the distribution of job types in the cleaned dataset. The dominance of Full-time positions, followed by Contract and Part-time roles, reflects the real-world structure of the LinkedIn job market, where full-time employment remains the standard offering. This distribution provides confidence that the dataset retains a representative sample of typical hiring practices rather than being biased toward niche or temporary roles.

Together, these diagnostics confirm that the cleaning and normalization procedures have produced a coherent, analysis-ready dataset with both completeness and structural balance, enabling the subsequent stages of skill, salary, and clustering analysis.

5. Market Signals in Job Roles, Skills, and Benefits (RQ1)

5.1. Overview and Objectives

This chapter addresses Research Question 1 (RQ1): What are the dominant market signals reflected in LinkedIn job postings in terms of job roles, skill requirements, and benefit offerings? The objective is to uncover which job categories, skills, and incentives define current labor market demand and how they collectively shape recruitment strategies across industries.

By analyzing job title frequencies, skill distributions, and benefit patterns, this chapter provides a descriptive foundation for understanding workforce composition and employer behavior in the digital labor economy. The analysis employs frequency aggregation, visualization, and comparative interpretation to identify key occupational clusters and incentive structures that make positions attractive to potential applicants.

5.2. Most Common Job Titles

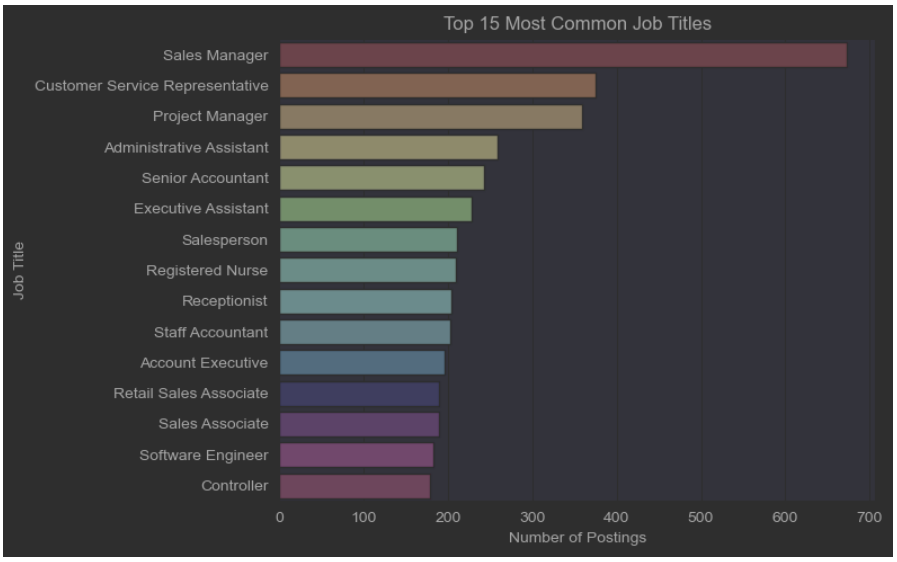

Figure.5. Top 15 Most Common Job Titles

The analysis begins with an exploration of job title frequencies to identify the most prevalent occupational categories in the LinkedIn dataset. As shown in Figure 5, the fifteen most common titles include Sales Manager, Customer Service Representative, and Project Manager. These roles dominate the dataset, accounting for a significant proportion of all job postings, and highlight the concentration of demand in management, client service, and coordination functions.

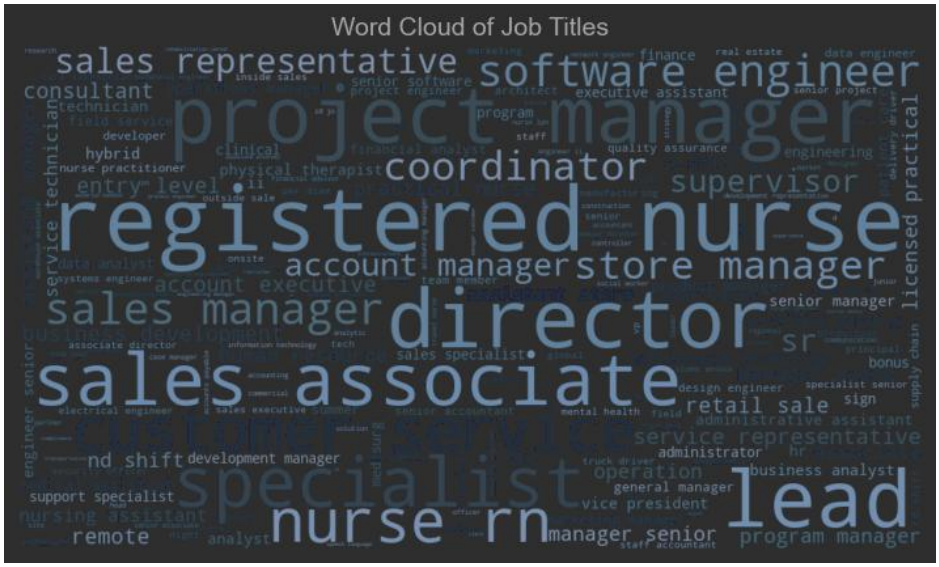

Figure.6. Word Cloud of Job Titles

Complementing this, Figure 6 presents a word cloud of all job titles, which provides a more holistic view by emphasizing the overall prominence of recurring terms. Frequent appearances of Registered Nurse, Director, Sales Associate, and Software Engineer reveal that while administrative and sales-oriented positions dominate numerically, substantial demand also exists in healthcare and technology domains.

The contrast between the bar chart and word cloud lies in analytical scope: the bar chart reflects discrete job title frequencies, while the word cloud aggregates similar expressions to reveal semantic prominence. Together, they illustrate that the LinkedIn job market features a dual focus: operational and managerial positions on one hand, and professional specializations such as engineering and healthcare on the other.

5.3. Top Required Skills

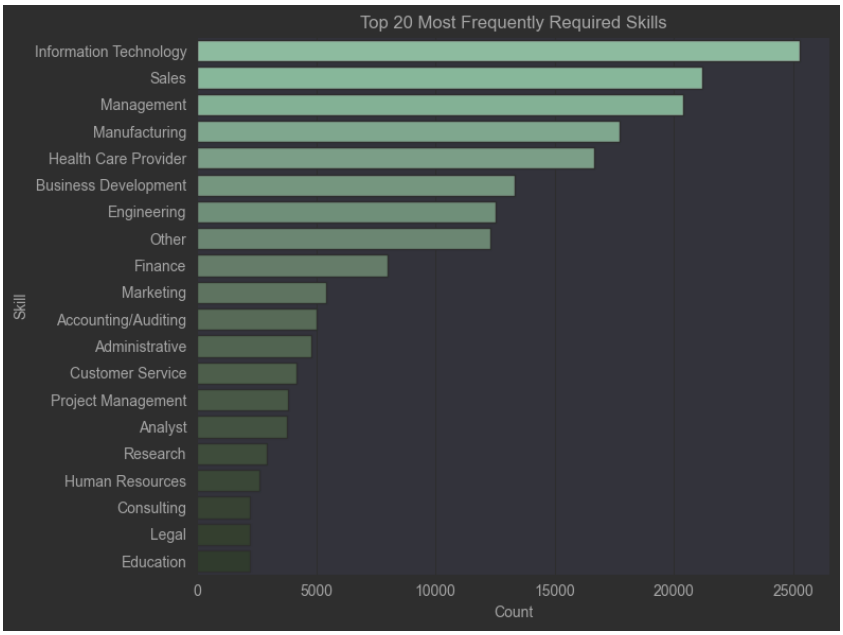

Figure.7. Top 20 Most Frequently Required Skills

The next analysis examines which skills are most frequently demanded across job

postings. Each posting's skill list was expanded and aggregated to compute total frequency per skill category. The results, shown in Figure 7, reveal that Information Technology, Sales, and Management are the three most in-demand domains.

This distribution demonstrates the dual-engine nature of today's labor market, where technological capabilities and business acumen jointly drive employability. Technical competencies such as IT, Engineering, and Manufacturing appear alongside leadership-oriented skills such as Management and Business Development, underscoring the increasing importance of cross-functional expertise in digitally transforming industries.

In addition, Health Care Provider appears prominently among the top skills, reflecting strong recruitment activity in essential service sectors. Stable demand is also observed for Finance, Marketing, and Accounting/Auditing, indicating that traditional business support functions continue to play a crucial role in corporate operations.

Overall, the skill landscape suggests that the modern job market rewards individuals who combine digital proficiency with strategic and managerial capabilities, implying that interdisciplinary professionals will maintain a sustained competitive advantage.

5.4. Skill Demand by Industry

To better understand how skill requirements differ across industries, the analysis cross-references industry categories with skill frequency counts. The resulting heatmap in Figure 8 visualizes the intensity of demand for each skill across major sectors.

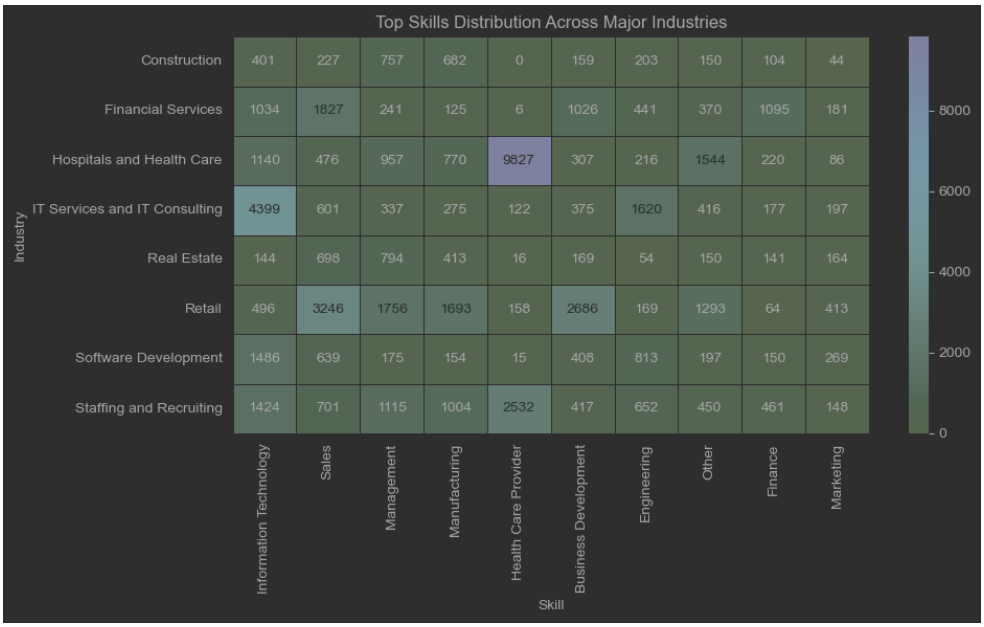

Figure.8. Top Skills Distribution Across Major Industries

Distinct specialization patterns emerge from this visualization. The IT Services and Consulting industry shows exceptionally high demand for Information Technology skills, while Hospitals and Health Care exhibit strong emphasis on healthcare-specific competencies. Similarly, Retail and Sales sectors concentrate on Sales and Management abilities, and Financial Services display higher frequency in Finance and Accounting skills, aligning with their business core.

The heatmap also demonstrates the pervasive presence of Management and Business Development across nearly all industries. This suggests that leadership, coordination, and organizational skills remain universally valued regardless of domain. Such results highlight

the hybridization of skill demand in the contemporary labor economy—technical depth is essential, yet it must coexist with generalist managerial strength.

5.5. Benefit Offerings and Market Attractiveness

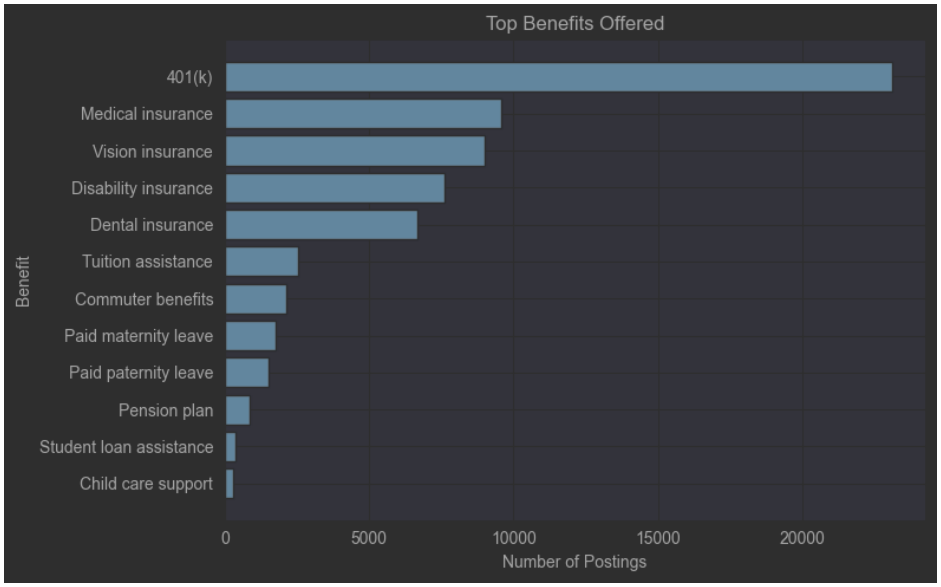

Figure.9. Top Benefits Offered

In addition to technical and managerial competencies, employers signal job attractiveness through the benefits they provide. Figure 9 summarizes the most frequently offered benefits across all postings. The 401(k) retirement plan ranks first, followed by Medical Insurance, Vision Insurance, and Dental Insurance, indicating that financial and health-related benefits remain the most common instruments for attracting and retaining talent in the U.S. job market.

Beyond these standard offerings, benefits such as Tuition Assistance and Commuter Benefits reflect employers' increasing focus on employee development and work–life integration. Meanwhile, provisions like Paid Maternity Leave, Paternity Leave, and Child Care Support demonstrate progressive adoption of family-oriented employment policies.

These patterns imply that firms are expanding beyond traditional compensation models to emphasize well-being, career growth, and flexibility as differentiating factors in talent acquisition. Benefits have thus evolved into a key strategic lever, conveying not only economic incentives but also organizational culture and employer branding.

5.6. Summary

The findings from this chapter collectively illustrate the structural dynamics of LinkedIn's job market. Sales, customer service, and project management positions dominate the landscape, while IT and healthcare remain essential technical anchors. The co-occurrence of business and digital skills demonstrates the increasing interdependence between commercial and technological expertise in modern employment ecosystems.

Moreover, the widespread inclusion of benefits such as retirement plans and healthcare insurance reflects both economic and cultural dimensions of job attractiveness. Employers use benefit design not only as a recruitment signal but also as a means of differentiation in competitive markets.

Together, these insights form the descriptive backbone of this study. They provide the empirical foundation for the next stage, which investigates salary transparency and sectoral

variation.

6. Salary Transparency and Disclosure Patterns (RQ2)

6.1. Overview and Objectives

This chapter addresses Research Question 2 (RQ2): How transparent are employers in disclosing salary information across different industries and work types, and what do these patterns reveal about organizational norms and labor market structures.

During the early data exploration stage, it became clear that building predictive models for salary estimation was not feasible. More than 90 percent of job postings lacked complete information in any of the three key salary fields (minimum, median, or maximum). Furthermore, inconsistent pay formats—such as a mix of hourly and annual wages—introduced normalization challenges.

Given these data limitations, the analysis in this chapter shifts focus from salary prediction to salary transparency. Instead of estimating compensation levels, it examines which industries and employment arrangements are more likely to disclose pay information. The objective is to use salary disclosure as a behavioral indicator of employer openness, compliance culture, and competitive positioning within the labor market.

6.2. Salary Disclosure Overview

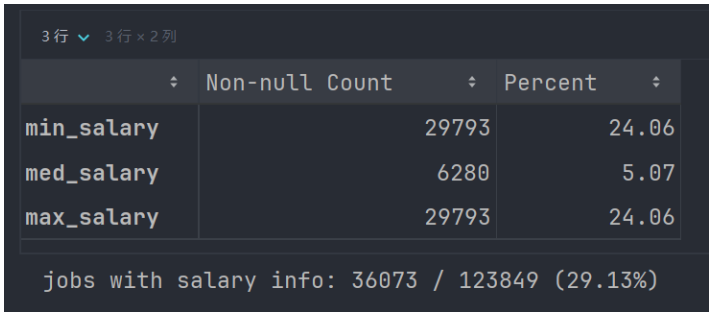

Figure.10. Completeness of salary-related fields and overall disclosure rate

The first step is to quantify the overall extent of salary disclosure in the dataset. As shown in Figure 10, only around 29.1% of all postings (36,073 out of 123,849) include at least one salary-related value. Among these, the minimum and maximum salary fields are populated in roughly 24% of records each, while the median salary field appears in just about 5%.

This finding indicates that salary disclosure remains the exception rather than the norm on LinkedIn job postings. The limited use of the median salary field further suggests that even among disclosing employers, most provide ranges rather than precise figures. The prevalence of incomplete or irregular pay data highlights continuing reluctance among organizations to publish compensation details, despite growing global advocacy for pay transparency.

6.3. Salary Disclosure by Industry

To understand how salary transparency differs across economic sectors, the dataset was expanded by decomposing multi-industry job tags so that each job–industry pair could be evaluated independently. This adjustment ensures that the disclosure rate computed for each industry accurately reflects its own reporting behavior rather than being diluted by

cross-sector overlaps.

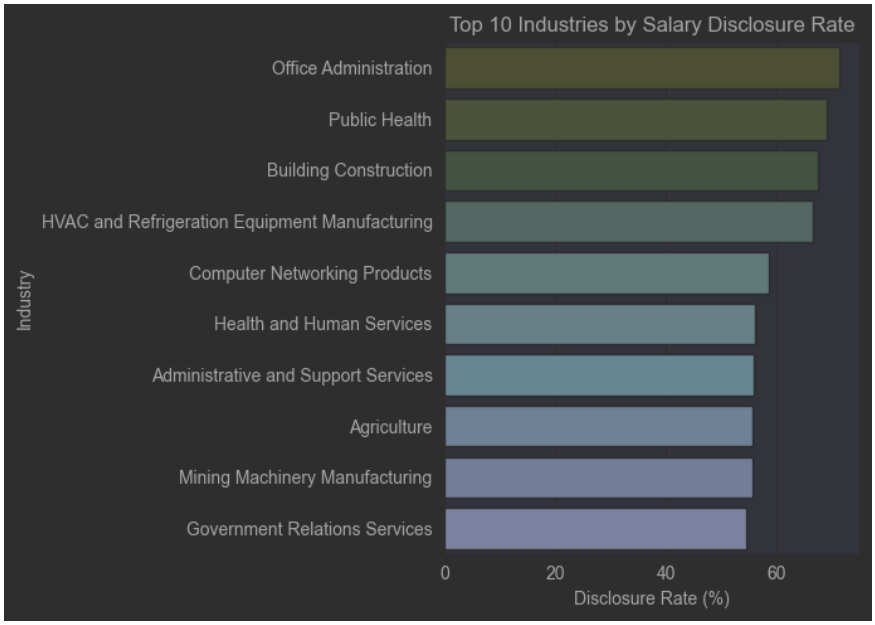

Figure.11. Top 10 Industries by Salary Disclosure Rate

Figure 11 presents the ten industries with the highest salary disclosure rates. The top performers include Office Administration, Public Health, and Building Construction, each exhibiting disclosure levels exceeding 65 percent. These sectors typically operate under standardized pay structures or formal public accountability frameworks, such as government procurement, healthcare systems, or unionized wage agreements. Within such environments, pay transparency is either mandated or culturally reinforced, making salary disclosure a normalized practice rather than a voluntary gesture. Other industries, such as HVAC and Refrigeration Manufacturing and Computer Networking Products, also demonstrate relatively strong transparency, likely reflecting their reliance on technical roles where compensation benchmarks are well established and competitively public.

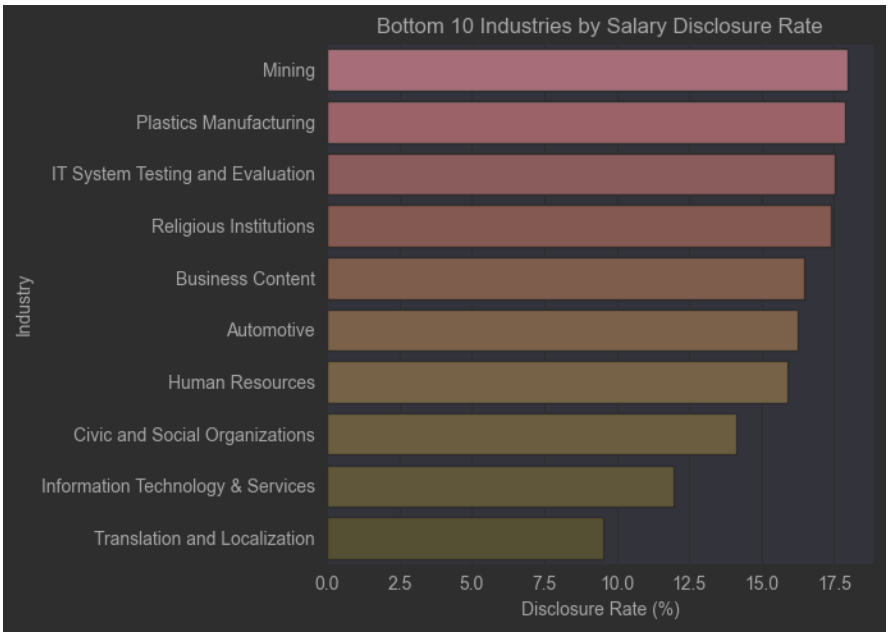

Figure.12. Bottom 10 Industries by Salary Disclosure Rate

In contrast, Figure 12 highlights the ten industries with the lowest levels of salary transparency. Mining, Plastics Manufacturing, and Religious Institutions appear at the bottom, with disclosure rates below 20 percent. These industries are characterized by high pay variability, limited regulatory oversight, or mission-driven operations, all of which reduce the incentive to disclose pay scales publicly. Similarly, Human Resources and Civic and Social Organizations show unexpectedly low disclosure, perhaps due to internal confidentiality norms or non-standardized remuneration structures. Even within the technology sector, Information Technology & Services and IT System Testing and Evaluation exhibit restrained transparency, suggesting that many private tech firms still treat compensation as proprietary information within a competitive market context.

Overall, the contrast between the two charts reveals a distinct polarization in salary disclosure behavior. Industries governed by public standards or formalized pay systems are more likely to publish compensation details, whereas sectors that rely on individualized negotiations or variable project-based contracts remain opaque. This divergence underscores the influence of institutional maturity, regulatory exposure, and market competition on the evolution of salary transparency.

6.4. Salary Disclosure by Work Type

In addition to industry segmentation, salary disclosure also varies by employment arrangement. The dataset categorizes jobs into multiple work types—such as full-time, part-time, temporary, and contract—allowing a comparison of disclosure norms across different contractual relationships.

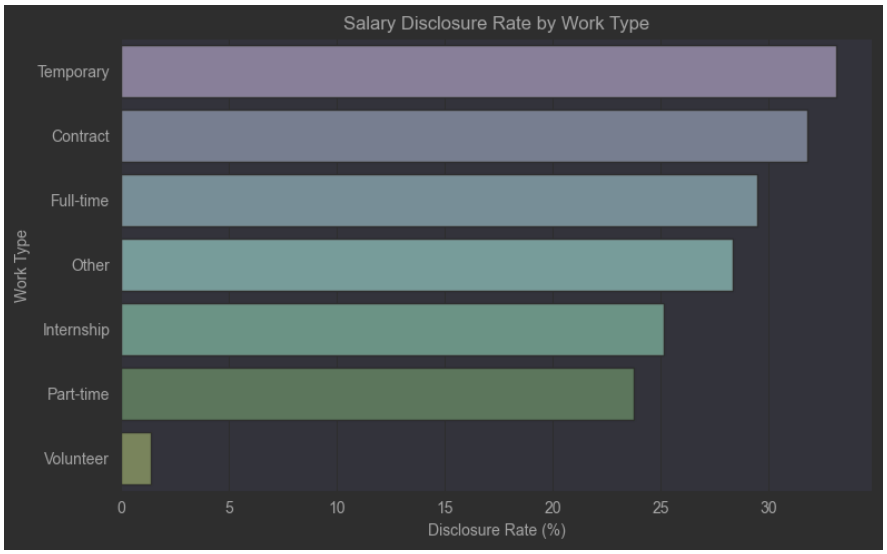

Figure.13. Salary disclosure rate by work type

As displayed in Figure 13, Temporary and Contract roles exhibit the highest transparency, with disclosure rates around 30%, surpassing even full-time positions. This finding aligns with the economic logic of short-term or project-based employment, where pay often serves as the primary basis for candidate evaluation. Because such positions typically lack long-term benefits, explicit compensation disclosure becomes a practical necessity for attracting applicants.

In contrast, Full-time and Other job types show moderate transparency (approximately 27–29%). These categories tend to operate under fixed salary bands or internal negotiation

processes, which reduces the perceived need for public disclosure.

Internship and Part-time postings demonstrate lower transparency (roughly 20–25%), consistent with their reliance on stipends or hourly wages that vary significantly by region and employer. Predictably, Volunteer positions record almost zero disclosure, reflecting their non-monetary nature.

These findings collectively suggest that salary transparency correlates not only with regulatory pressure and industry norms but also with the economic structure of employment contracts—the shorter and more transactional the arrangement, the more likely compensation is to be stated explicitly.

6.5. Summary

The analysis of salary disclosure reveals that transparency remains uneven across both industries and employment types. The overall dataset average of 29% indicates that most organizations still view compensation as confidential information, despite ongoing global trends advocating openness in pay communication.

Industries with standardized wage structures or public oversight—such as healthcare, construction, and administration—tend to disclose pay more frequently, while private, resource-based, and mission-driven sectors remain opaque. Similarly, work arrangements that rely on short-term contracts or gig-style labor display higher transparency than traditional full-time employment, suggesting a growing divide between transactional openness and institutional discretion in labor markets.

From a policy perspective, these results highlight the importance of legislative and cultural incentives in driving pay transparency. As more jurisdictions adopt mandatory disclosure laws, variations like those observed here may diminish. For now, however, transparency serves as a useful proxy for understanding organizational priorities, market competition, and trust dynamics between employers and job seekers.

This chapter thus provides a crucial transition from descriptive market characteristics to evaluative insights on compensation behavior. The next analysis builds upon this foundation by applying text-mining and clustering techniques to uncover latent job categories and structural patterns hidden within job descriptions.

7. Uncovering Hidden Job Categories through Clustering (RQ3)

7.1. Overview and Objectives

While the previous chapters focused on descriptive market characteristics and salary transparency, this chapter addresses Research Question 3 (RQ3): Can unsupervised text-based clustering uncover latent occupational structures that are not explicitly represented by job titles or industry labels?

Traditional categories such as “industry” or “job title” are often too coarse to capture the thematic nuances of modern employment. Many postings combine business and technical functions, creating hybrid roles (e.g., data-driven product manager or technical marketing analyst). To detect such hidden structures, this study applies a text-mining and clustering pipeline based on job descriptions, titles, and skills. The approach uses TF-IDF vectorization to quantify textual importance, Truncated SVD to compress high-dimensional representations, and K-Means clustering to group semantically similar postings.

This unsupervised framework offers a data-driven way to reveal emergent job domains that extend beyond predefined occupational taxonomies, helping identify meaningful clusters that reflect the evolving digital labor landscape.

7.2. Data Preparation and TF-IDF Vectorization

Each job posting was represented by merging textual fields—title, description, and skill names—into a single document-level text. Common non-informative terms were removed, and words were normalized to lowercase.

A TF-IDF vectorizer was then applied to transform text into numerical features, assigning higher weights to terms that are distinctive within the corpus while discounting globally frequent ones. The vocabulary was limited to 5 000 terms, with thresholds of min_df = 5 and max_df = 0.8 to filter out rare or overly common tokens. This ensured that subsequent clustering captured meaningful semantic distinctions rather than noise.

7.3. Text Representation and Dimensionality Reduction

Although TF-IDF provides rich lexical information, the resulting feature space is sparse and computationally expensive. To address this, Truncated SVD (a linear form of latent semantic analysis) was used to reduce dimensionality to 50 components. This transformation projects postings into a latent semantic space, preserving the main variance directions while smoothing minor lexical variations.

```
tf-idf shape: (123849, 5000)
reduced shape: (123849, 50)
```

Figure.14. Reduced shape

The reduced representation enables efficient clustering and enhances interpretability by summarizing semantic similarity across occupations.

7.4. Clustering and Keyword Extraction

To group semantically related postings, the reduced vectors were clustered using K-Means. The optimal number of clusters (k) was determined via Elbow (inertia) and Silhouette score analysis.

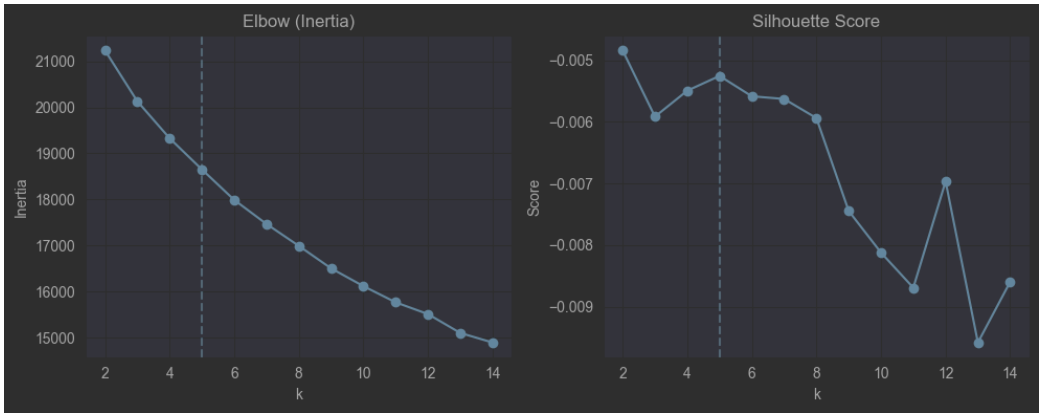

Figure.15. Elbow and Silhouette Analysis for Optimal k

Figure 15 shows both evaluation curves. The Elbow method displays a clear inflection at k = 5, beyond which inertia decreases more slowly. The Silhouette score, which measures cluster compactness and separation, also peaks near k = 5. Therefore, k = 5 was selected as the most balanced configuration for the job-level dataset.

Following clustering, the centroid vectors were mapped back to the original TF-IDF space to extract the 12 most representative keywords per cluster. These keywords summarize the primary thematic focus of each occupational group.

To visualize the semantic structure, PCA was used to project the 50-dimensional embeddings into two principal components.

Figure.16.2D Visualization of Job Clusters

Figure 16 shows the resulting 2D distribution: postings form five dense but partially overlapping regions.

Distinct clusters correspond to operational, retail, healthcare, business, and technical domains, while overlaps indicate hybrid job categories that span multiple functions (e.g., data-driven marketing or engineering-management roles).

7.5. Summary

The unsupervised clustering analysis successfully identified five coherent occupational groups derived purely from the textual semantics of job postings.

Cluster	Example Top Keywords	Representative Theme
0	work, experience, job, service, required, company, team, ability, equipment, skills	Operational and Support Services
1	store, sales, customer, retail, merchandise, manager, associate	Retail and Customer Interaction
2	care, patient, nursing, health, medical, clinical, provider	Healthcare and Nursing
3	sales, marketing, business, customer, development, product, client	Business and Marketing
4	project, data, engineering, design, management, software, technical	Engineering and Technical Functions

Table.1. Top Keywords and Inferred Themes per Cluster

By applying TF-IDF representation, Truncated SVD dimensionality reduction, and K-Means clustering (k = 5), the model uncovered distinct job categories that extend beyond

conventional industry or title-based classifications.

The clustering results reveal five major occupational themes that align closely with labor-market structures observed in earlier descriptive analyses:

1) **Cluster 0 – Operational and Support Services:**

Broadly defined service and logistics roles emphasizing task execution, teamwork, and basic skills. These postings frequently appear across manufacturing, maintenance, and administrative functions.

2) **Cluster 1 – Retail and Customer Interaction:**

Sales associates, store managers, and frontline service roles that rely heavily on interpersonal skills and direct consumer contact. The prominence of this cluster reinforces the continuing demand for retail and service operations in the post-pandemic economy.

3) **Cluster 2 – Healthcare and Nursing:**

Distinctive lexical concentration around "patient," "nursing," and "clinical," indicating clear specialization. This group captures medical assistants, registered nurses, and other health service professionals.

4) **Cluster 3 – Business and Marketing:**

Roles oriented toward organizational growth and client relations, including sales, marketing strategy, and business development. This reflects firms' emphasis on customer acquisition and digital marketing integration.

5) **Cluster 4 – Engineering and Technical Functions:**

Highly technical positions involving data analysis, software development, and project management. These roles form the analytical backbone of digitally transformed industries.

The overlap observed in the PCA plot suggests that modern job boundaries are increasingly fluid. Hybrid positions occupy transitional regions between clusters. This finding underscores the emergence of cross-domain competence as a key feature of the digital labor market.

By integrating text vectorization, dimensionality reduction, and unsupervised clustering, this chapter demonstrates that latent occupational patterns can be identified directly from textual data without manual labels. The five discovered clusters mirror core employment domains while highlighting emerging hybrid roles that defy traditional industry boundaries. These insights lay the foundation for Chapter 8, which examines how specific skills and skill combinations relate to salary outcomes and market value.

8. Skill Premium and Salary Gap Analysis (RQ4)

8.1. Overview and Objectives

Building on the structural insights from the previous chapter, this section addresses Research Question 4 (RQ4): Do specific skills or combinations of skills lead to measurable salary advantages in the digital labor market?

While earlier analyses (RQ1–RQ3) examined market signals and job clustering, this chapter focuses on the economic dimension of skills—specifically, how skill type, popularity, and combination patterns influence compensation outcomes. The analysis aims to quantify the skill premium, assess whether market popularity translates into pay advantage, and identify cross-domain skill pairings that produce salary uplift effects.

8.2. Data Preparation and Methodology

Only job postings with at least one valid salary field (min_salary, med_salary, or max_salary) were included, yielding approximately 36,000 records—roughly 29% of the full dataset.

The skill_names field was tokenized and exploded to create one record per (job, skill) pair. Subsequently, mean and median salaries were aggregated by skill to derive single-skill statistics. Skills with fewer than 50 associated postings were removed to ensure statistical reliability. All salaries were normalized to U.S. dollars for comparability.

8.3. Top-Paying Skills

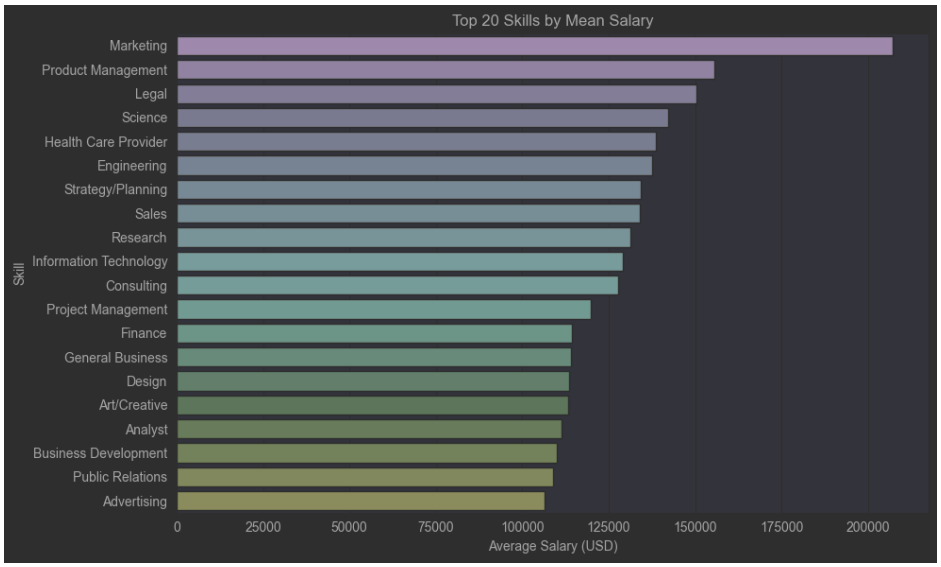

Figure.17. Top 20 Skills by Mean Salary

Figure 17 shows the 20 highest-paying skills ranked by mean salary. Marketing, Product Management, and Legal stand out as the top three skills, each commanding average salaries above USD 150,000, with Marketing exceeding USD 200,000. These domains typically involve strategic decision-making, leadership, and revenue-generating responsibility, explaining their high compensation. From a technical standpoint, Engineering and Information Technology maintain strong positions, averaging USD 130,000–140,000, confirming persistent demand for technical talent.

However, their pay levels now approach those of business-oriented functions, suggesting a convergence between technical specialization and managerial influence. Creative and support-oriented skills—such as Design, Public Relations, and Advertising—occupy the lower tier, averaging around USD 100,000. These roles remain vital for brand identity and communication but lack the high leverage or strategic control of leadership and engineering positions.

Overall, the data demonstrate a pronounced skill premium effect: specialized and cross-functional skills are rewarded disproportionately higher, particularly where they connect business strategy and technical execution.

8.4. Skill Popularity vs. Compensation

To test whether frequently requested skills also yield higher salaries, a log-scale regression was fitted between skill frequency and average salary. A Pearson correlation analysis quantified the relationship. The resulting correlation coefficient ($r = 0.202$, $p =$

0.244) indicates a weak and statistically insignificant association between skill popularity and pay.

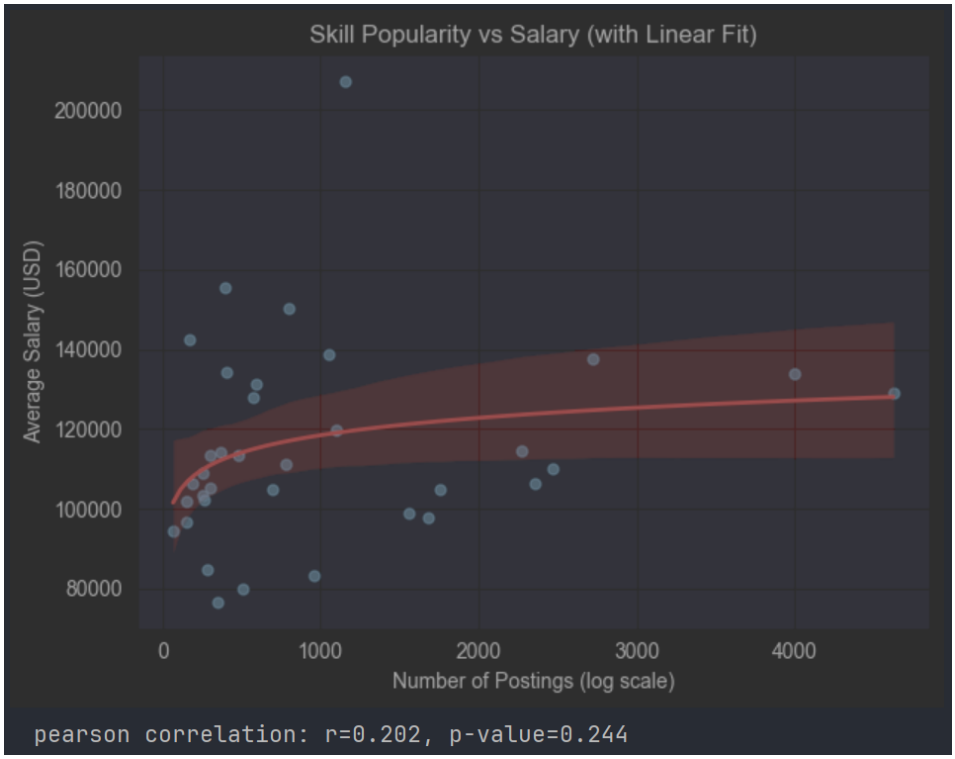

Figure.18. Skill Popularity vs. Salary with Linear Fit

Although the regression line shows a slight positive slope, the high p-value implies that popularity alone does not predict salary level. This result suggests that high-frequency skills—such as communication, teamwork, or customer service—do not necessarily command higher compensation. Instead, rarity and specialization play a stronger role: less common skills (e.g., legal, data science, product management) yield higher salaries despite appearing in fewer postings. Hence, in the digital labor economy, skill scarcity outweighs skill ubiquity as a determinant of pay.\

8.5. Cross-Skill Combinations and Salary Uplift

Beyond individual skills, the analysis explores whether multi-skill combinations produce synergistic pay advantages. For each job, all unique skill pairs were generated, and the average salary of each pair was compared against the baseline mean of its two component skills. The uplift metric is defined as:

$$\text{Uplift}(A, B) = \bar{S}_{A,B} - \frac{\bar{S}_A + \bar{S}_B}{2}$$

where $\bar{S}_{A,B}$ represents the mean salary for postings containing both skills. Pairs with at least 30 co-occurrences were retained for reliability.

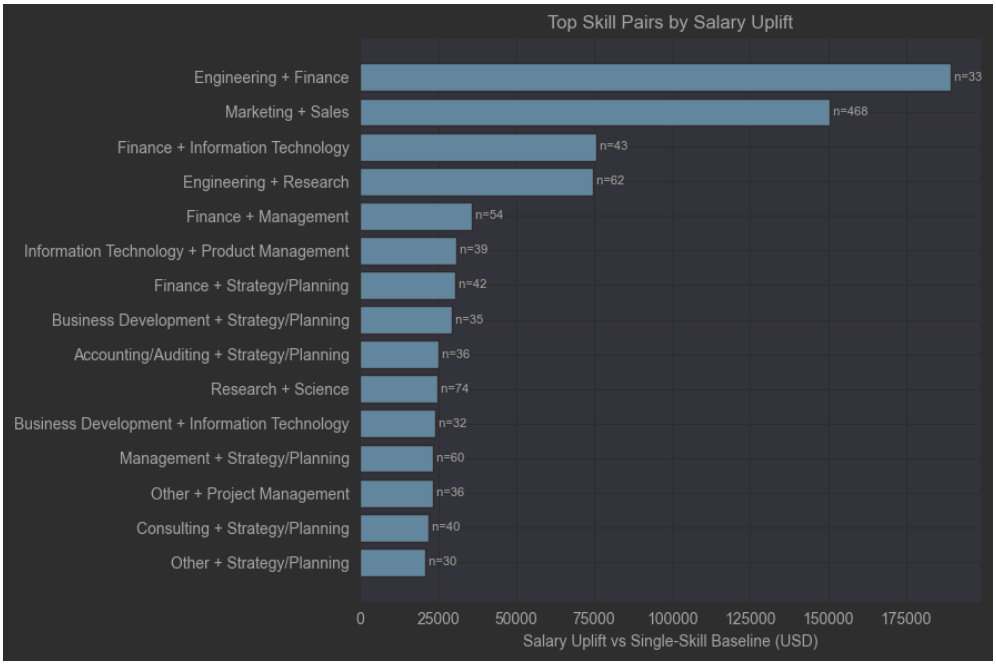

Figure.19.Top Skill Pairs by Salary Uplift

Figure 19 illustrates the top uplift pairs. The leading combinations—Engineering + Finance, Marketing + Sales, and Finance + Information Technology—display salary uplifts between USD 120,000–180,000 above the baseline. Interestingly, Engineering + Finance appears among the highest uplift pairs despite relatively small sample size (n = 33), implying it represents a niche yet highly valued intersection between analytical and business capabilities.

Across all pairs, a consistent pattern emerges:

Combinations bridging technical and business domains (e.g., Finance + IT, Engineering + Management) consistently yield high returns.

Strategy-related skills (Strategy/Planning, Product Management) frequently appear in top uplift pairs, reinforcing their role as connectors between operational and executive functions.

Purely creative or operational combinations show limited uplift, suggesting that market premiums accrue primarily to cross-domain integration skills.

8.6. Summary

The findings from RQ4 confirm the existence of a measurable skill premium in the digital labor market, structured by both specialization and interdisciplinarity.

Specialization Premium: Strategic and professional skills (Marketing, Product Management, Legal, Engineering) exhibit the highest average salaries, reflecting concentrated demand for high-responsibility roles.

Scarcity Effect: Skill frequency shows no strong correlation with salary, implying that market scarcity rather than popularity drives compensation.

Synergy Advantage: Skill pairs combining technical and business competencies produce significant salary uplift, underscoring the market's preference for hybrid, boundary-spanning professionals.

Together, these results highlight that economic value in the digital workforce increasingly depends on the integration of diverse knowledge domains. Professionals

capable of combining analytical, technical, and managerial skills occupy a strategic position in the labor hierarchy—representing the core archetype of Industry 4.0 talent.

9. Conclusion

9.1. Summary of Key Findings

This study provides an integrated view of the digital labor market through four analytical dimensions.

The results reveal that job postings on LinkedIn reflect a market dominated by managerial, sales, and administrative roles, with technical and analytical competencies—particularly in information technology and engineering—remaining central to demand. Benefits such as health insurance and retirement plans also serve as key market signals that enhance job attractiveness.

The examination of salary disclosure patterns shows that only about one-third of employers provide pay information, with transparency varying considerably across industries and work arrangements. Public-facing and regulated sectors, such as healthcare and construction, demonstrate higher disclosure rates, whereas private and technology-oriented fields tend to remain opaque.

Unsupervised text clustering further uncovered latent occupational structures, identifying five coherent job domains—operations, retail, healthcare, business, and technology. These categories highlight an increasingly hybrid labor environment where traditional functional boundaries are blurred by digital transformation.

Finally, the analysis of skill–salary relationships confirmed the existence of a distinct skill premium. Specialized and cross-functional capabilities, especially those bridging technical and business expertise, command substantially higher pay. While skill popularity does not correlate strongly with compensation, rare and interdisciplinary skill sets—such as engineering combined with finance or product management—offer the greatest salary advantages.

Collectively, these findings suggest that digital-era employability depends not only on technical proficiency but also on the capacity to integrate strategic, analytical, and managerial knowledge within a rapidly evolving labor ecosystem.

9.2. Limitations and Methodological Constraints

Despite the analytical depth of this study, several limitations should be acknowledged. First, the dataset is constrained by incomplete salary information—over two-thirds of postings lack explicit pay data—limiting the scope of quantitative modeling. Second, the LinkedIn sample may overrepresent white-collar and technology-oriented roles, resulting in partial coverage of the broader labor market. Third, textual heterogeneity in job titles and skill descriptions introduces potential noise, even after standardization and cleaning. Additionally, the study captures a single temporal snapshot, whereas market dynamics and skill valuations evolve continuously.

Finally, salary normalization across currencies and pay periods remains an approximation, potentially obscuring fine-grained differences. These factors do not undermine the validity of the findings but highlight the interpretive boundaries within which they should be understood.

9.3. Future Research Directions

Future research could build on this work in several ways. Longitudinal datasets from multiple platforms would allow the tracking of temporal shifts in job demand and pay transparency. Integrating contextual embeddings or large language models could enhance the semantic precision of clustering and skill extraction. Further, incorporating firm-level and regional indicators would help explain variations driven by geography, regulation, or company scale.

Finally, extending the analysis toward fairness and equity: examining gender, experience, or diversity dimensions, would contribute to a more inclusive understanding of digital labor dynamics.

References

[1]. Arsh Koneru (2023). LinkedIn Job Postings Dataset (2023–2024). Kaggle. https://www.kaggle.com/datasets/arshkon/linkedin-job-postings

[2]. GeeksforGeeks. Understanding TF-IDF (Term Frequency-Inverse Document Frequency) . https://www.geeksforgeeks.org/machine-learning/understanding-tf-idf-term-frequency-inverse-document-frequency/

[3]. Scikit-learn Developers. TruncatedSVD — scikit-learn 1.5.0 documentation. https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.TruncatedSVD.html

[4]. Analytics Vidhya. K-Means Clustering Algorithm. https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/

[5]. Scribbr. Pearson Correlation Coefficient (r) | Guide & Examples. https://www.scribbr.com/statistics/pearson-correlation-coefficient/

Appendix

Table.1. AI Usage Statement

AI Tool Used	Prompt and Output	How the Output Was Used
ChatGPT	Diagnosed and explained the FutureWarning in seaborn: Passing palette without assigning hue is deprecated and will be removed in v0.14.0. Suggested replacing the original call with sns.barplot(x=..., y=..., hue='y', legend=False, palette='Blues_r').	Implemented the revised syntax in visualization code to maintain forward compatibility with seaborn ⩾0.14.0 and suppress the warning during execution.
ChatGPT	Suggested parameter tuning for TF-IDF, TruncatedSVD, and K-Means to optimize cluster interpretability and reduce sparsity.	Adopted recommended configurations for consistency and reproducibility.
ChatGPT	Helped formulate code for salary-related aggregation and skill-pair uplift computation using itertools.combinations and Pandas group operations.	Implemented the salary uplift formula and applied filters based on AI suggestions to ensure statistical robustness.
ChatGPT	Consulted on integrating multiple relational tables (job_postings, skills, companies) containing overlapping keys (company_id). AI proposed a hierarchical merge strategy: first link job → company, then outer-join with skills and benefits.	Implemented the unified data cleaning pipeline based on this hierarchy, reducing redundancy and ensuring referential integrity across tables.
ChatGPT	Discussed KMeans Elbow curve issue where no clear inflection point appeared due to high text dimensionality. AI suggested dimensionality reduction via TF-IDF top 5000 features before retraining.	Adopted TF-IDF feature capping and re-ran clustering, resulting in improved silhouette scores and clearer cluster separability.

ChatGPT	Diagnosed mis-splitting of composite industry labels such as "Oil, Gas and Mining" into three separate entries. Suggested temporarily replacing commas with ampersands ("&") as placeholders before splitting and later restoring.	Revised the preprocessing step to preserve compound industry names, ensuring accurate aggregation in industry-level analyses.
ChatGPT	Helped resolve InvalidParameterError: stop_words caused by passing a frozenset to TfidfVectorizer. Recommended converting it to list(custom_stops) and refining token pattern to r"(?u)\b[a-zA-Z][a-zA-Z+\-#\.]{1,}\b".	Updated the vectorizer initialization code accordingly, eliminating the error and improving token filtering by restricting to valid English terms.
DeepL	Translated Chinese sections of the report and figure explanations into English for submission.	Used translations as preliminary drafts; manually edited for accuracy and fluency in the final version.
SimpleTex	Converted mathematical expressions from markdown to LaTeX format for clean inclusion in the report and slides.	Integrated exported LaTeX equations into Word using MathType for consistent notation and formatting.

Table.2. Sample Record from Cleaned Postings Dataset

Field Name	Description / Example
job_id	921716
company_id	2774458
title	*Marketing Coordinator*
description	"A leading real estate firm in New Jersey is seeking an administrative Marketing Coordinator with experience in graphic design. You will be working closely with our sales and executive teams in a fast-paced environment. … Responsible for preparing print materials, managing event vendors, fulfilling agent design requests, and maintaining brand assets." *(excerpt)*
location	New Jersey, United States
formatted_work_type	Full-time

remote_flag_simple	False
formatted_experience_level	Entry-Level
skill_names	Adobe Creative Cloud; Illustrator; Photoshop; Graphic Design; Marketing Coordination
industry_names	Real Estate; Marketing & Advertising
benefits	Paid Time Off; Flexible Schedule
pay_period	Hourly
currency	USD
min_salary	18
max_salary	20
avg_salary	19
company_name	Confidential Real Estate Firm
company_size	51–200 employees
country	United States
city	New Jersey
employee_count	120
follower_count	1,320
posting_domain	linkedin.com
application_url	https://www.linkedin.com/jobs/view/921716/
applies	48
views	1,230
remote_allowed	False
time_recorded	2024-03-21 12:35:00

Note:

This sample record is extracted from the cleaned_postings.csv dataset after preprocessing and integration. It demonstrates the normalized structure of job-related, company-related, and salary-related fields retained for analysis. Long text fields are abbreviated for readability.

In addition to the LinkedIn job postings analysis presented in the main report, other members of the project team also conducted independent research, including an extended exploration using a separate resume dataset https://www.kaggle.com/datasets/saugataroyarghya/resume-dataset .

These supplementary studies covered topics such as skill extraction, clustering analysis, and correlations between education and work experience. Given the limited time and the complexity of integrating different analytical pipelines, these results were not incorporated into the main project but are summarized here as reference materials to illustrate the broader scope of the group's collective efforts.

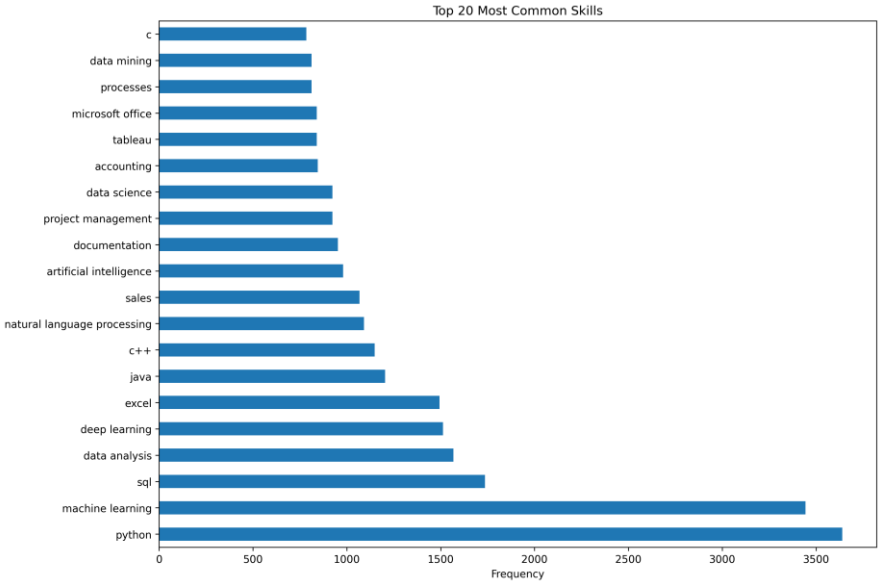

Figure.1. Top 20 Most Common Skills

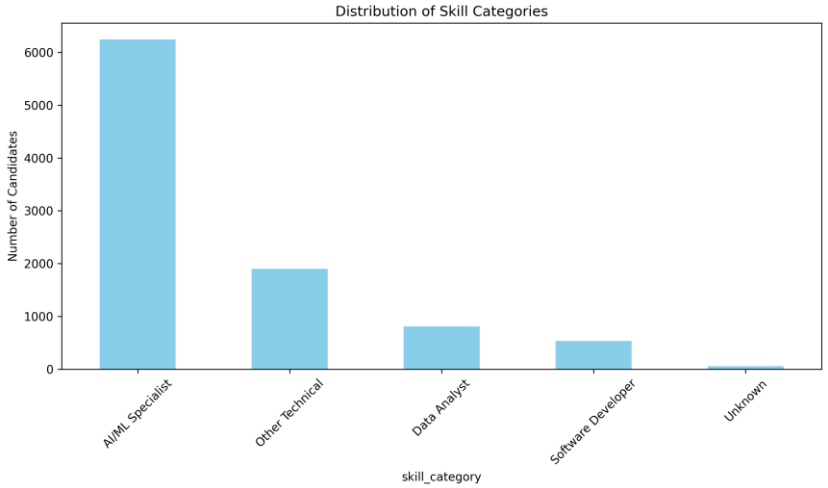

Figure.2. Distribution of Skill Categories

Figure.3. Word Cloud for skills (postings dataset)

Figure.4. Word Cloud for skills (resume dataset)

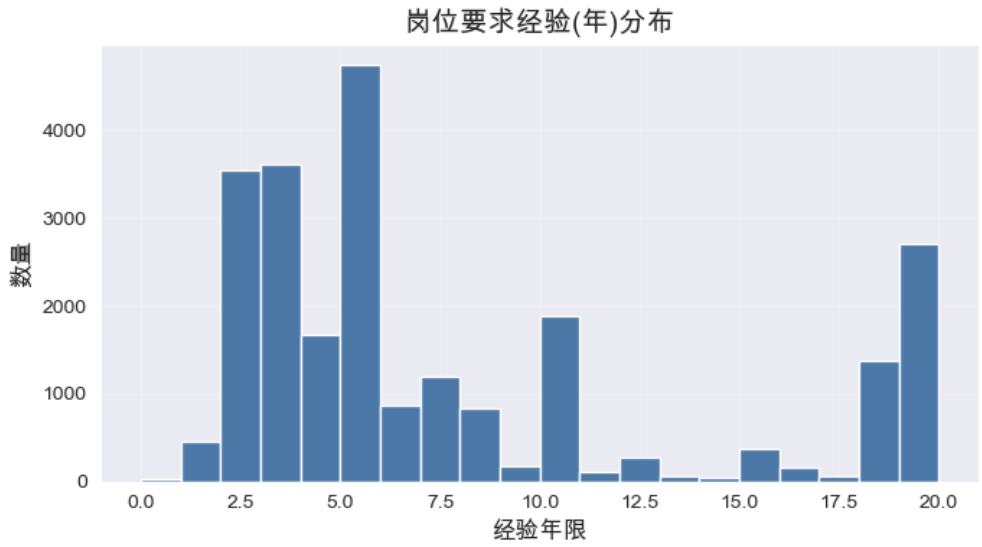

Figure.5. Distribution of required years of experience