

Career Navigation with LinkedIn Job Posting Insights

What skills, careers, and cities should you be working towards?

Alex Archer

Applied Computer Science
University of Colorado,
Boulder
Boulder, Colorado, USA
alar2300@colorado.edu

Project Statement

In today's rapidly evolving job market, finding the right career path can be like searching for a needle in a haystack. With industries constantly transforming and demanding new skills at record speeds, job seekers and professionals are often left wondering, 'Where do I fit in?' Which my project, "Career Navigation with LinkedIn Job Posting Insights," I intend to help answer this question. With the help of the massive public dataset "1.3 M LinkedIn Jobs & Skills (2024)," I aim to provide clarity about the current job market landscape, turning this massive data set into relevant and up-to-date insights.

This dataset will be a source of useful insights into current opportunities, with 1.3 million entries spanning numerous industries. Each job listing provides abundant information, detailing not only the role, but the skills it demands, which company is offering, and the location of the posting. The attributes in the first of the two tables being examined include the link to the posting as the primary key, job title, company, job location, and search city, which is the criterion for collecting location-relevant postings. The second table holds the list of relevant skills for each posting, as well as the postings link as the primary key.

My goal is twofold. First, I will discover where jobs in specific industries are booming, so that job seekers know where to find opportunities best suited to their skill set. Second, I will uncover the most sought-after job titles in addition to the top skills they require by region, thus providing a roadmap for skill development and career planning. I hope to reveal the current state of the 2024 job market, helping to answer questions such as, 'In what cities are my skills in demand?', 'What job titles match my current

skillset?', 'what skills do I need to develop for my desired job title?'

What I find most interesting, and will try to reveal the patterns of, is the complex interlacing between job titles and skills within industries. What are the niche skills that can help broaden the scope of your job search? Are there emerging or established hotspots for specific industries? The knowledge mined from this dataset is not only academic but also immensely practical. It could be useful for guiding educational institutions in relevant curriculum development, and also helping job seekers choose in-demand skills to attain.

In essence, my project seeks to make it easy to make informed choices about your career, powered by data-driven insights from the heart of the professional world: LinkedIn.

Part 3: Updated Motivation

Navigating the evolving climate of the LinkedIn job market, my project, "*Career Navigation with LinkedIn Job Posting Insights*," has made significant strides toward clarifying the complexities of job and skill alignment in 2024. The initial motivation for this project was fueled by the challenge many including myself face in pinpointing career opportunities that align best with their skills in a rapidly changing environment. Utilizing the "1.3 M LinkedIn Jobs & Skills (2024)" dataset has not only reinforced the importance of this analysis but also shows the practical pathways this project can provide for job seekers, educational bodies, and industry professionals.

Since the beginning of this project, the collected data has been shining a light on the current and emerging market demands. The process of cleaning, validating,

and structuring over a million job postings has provided a more granular view of the job landscape that was previously hidden by the sheer scale and disorganization typical of such large datasets. By categorizing job titles into broad industry categories and linking them to specific skills, I am beginning to uncover critical trends that will guide individuals in making more informed career decisions.

The insights derived so far are promising. They show not only where jobs specific to certain industries are booming but also detail the skills that are becoming increasingly crucial in these industry categories. This aligns perfectly with our goal to equip job seekers with the knowledge of where their skills are needed the most and what new skills they might need to acquire. Similarly, for educational institutions, these insights are important for aligning lesson plans with market-relevant skills, ensuring that the next generation of professionals is ready for the future job market.

As I continue to refine the analysis and delve deeper into this dataset, the patterns I discover will become even more insightful, providing a clearer roadmap for navigating the current professional landscape.

Literature survey

The first piece of literature that I surveyed in preparation for this project is entitled, "Salience and Market-aware Skill Extraction for Job Targeting," which is a research paper directly from engineers at LinkedIn that elaborates on systems utilized by LinkedIn for skill extraction. This LinkedIn system addresses the limitations of traditional text-based approaches which often overlook skill salience (importance) and market dynamics. With the integration of market supply and demand factors, the method utilized significantly enhances job recommendation and skill suggestion accuracy, as evidenced by improved application rates (+1.92% job application rate) and reduced suggestion rejection rates (-37% rejection rate). Through detailed case studies, this paper offers essential insights into how these techniques vary across different industries, countries, and skill levels, demonstrating the practical benefits of their market-aware skill extraction system. These systems inform the inherent value of my own

skill extraction and the accuracy of the insights I intend to attain.

The second piece of literature that I analyzed in preparation for my project is an open-source Jupyter Notebook entitled, "Analysis of 1.3 Million Jobs on LinkedIn," authored by Galal Qassas. This notebook provides a detailed exploration of the skill demands reflected in LinkedIn and the 1.3 million job postings held within the dataset. By employing Python libraries including numpy, pandas, matplotlib, and seaborn, the notebook systematically parses and analyzes the dataset to identify prevalent skills across job descriptions. Notably, the analysis highlights the top 15 skills in demand, including communication, teamwork, and customer service, and uncovers that an average of 19 specific skills is typically required to apply for jobs on LinkedIn. This analysis is highly relevant to my project given how it offers a direct glimpse into the skills landscape, aiding in my foundational understanding of LinkedIn skill trends as well as the most common demands within the job market as a whole, thus aligning closely with my objective of mapping career paths and skill requirements by industry.

Lastly, the open-source Jupyter Notebook entitled, 'EDA on LinkedIn Jobs and Skills' conducts a comprehensive exploratory data analysis using Python and libraries such as numpy, pandas, matplotlib, and seaborn. It first confirms the dataset's integrity by ensuring that there are no duplicate entries and that all columns are non-null. The analysis further explores the dataset to discover some surface-level insights including top job titles, leading companies by job postings, and primary job locations, highlighting cities and regions with the most postings. Additionally, it counts the number of jobs at associate v.s. mid-level and senior positions, and position settings including remote, hybrid, or onsite positions which give a detailed overview of the dataset as a whole.

Proposed Work

To grant clarity in the 2024 job market landscape and provide valuable insights, my project begins in the data collection phase which will leverage the "1.3 M LinkedIn Jobs & Skills (2024)" dataset. The initial step in this process involves data cleaning to remove listings either with incomplete information or those

lacking in job summaries or skill details. Following this initial step, I will implement a validation process that ensures the integrity of a selected subset of job postings through manual checks.

My data preprocessing stage will integrate job postings with their corresponding skills given that the source data formats this information in separate tables, transforming this data into a structured format most convenient to my analysis. Job titles will then be categorized into broad industry sectors, with a new 'industry' column created for better classification. This structured data will form the foundation for a two-sided approach.

First, I will conduct a time-series analysis to identify and visualize the job market trends for 2024, with a focus on regional industry booms and the popularity of specific job titles within those regions and industries. This analysis aims to pinpoint where specific industries are thriving geographically, answering the question of where job opportunities are emerging most heavily.

Second, I will utilize clustering techniques, with k-means as the prime candidate, to group similar job titles and associated skills within industry sectors. This step is designed to uncover the patterns in skill demand, providing insights into the most sought-after skills by region and job category. This analysis hopes to reveal skills that can broaden job search scopes and identify both emerging and established industry hotspots.

My project sets itself apart from others both on this subject and on this dataset by not only mapping the current state of the job market extensively but also providing a forward-thinking perspective on skill demand trends and job opportunities. My current approach aims to arm job seekers as well as educational institutions and industry professionals with the knowledge to make more informed decisions, helping to bridge the gap between skill development and market needs.

Part 3: Updated Proposed Work

Building on the foundation laid out in the original project plan, significant progress has been made in clarifying the 2024 job market landscape using the "1.3 M LinkedIn Jobs & Skills (2024)" dataset. After the initial data collection phase, I successfully

completed extensive data cleaning to eliminate incomplete or inadequate job listings. This essential first step ensured that the analysis would be based on reliable and comprehensive data.

The subsequent data preprocessing work involved a process of integrating job postings with their corresponding skills from separate data tables and transforming this list into a form better suiting the analysis of individual skills. This step was crucial in transforming data into a structured, analyzable format. I introduced a new 'industry' column, enhancing the ability to categorize job titles into broad industry sectors accurately. This restructured data now serves as the backbone for my approach.

I created and refined a keyword map, correlating keywords in job titles to a manually refined map of industry-specific keywords. This initial approach proved to be slightly inaccurate due to ambiguous job titles. Industry classification was further refined and validated by comparing the skills list of each job posting to a curated list of the top 30 skills in the industry to which it was assigned. If the posting had any skills in common with the top 30 for its industry it was deemed a valid categorization, and invalid otherwise. This form of validation removed over 200,000 ambiguous classifications.

The second phase of the proposed work still involves applying advanced clustering techniques, specifically k-means clustering, to group similar job titles and associated skills within identified industry sectors. This methodology has the goal of exposing the underlying patterns in skill demands, highlighting the skills that are most valuable in various regions and job categories. Through this analysis, I expect to identify emerging skills and established industry hotspots, which can significantly widen job search horizons and aid career planning.

As the project moves forward, the focus will continue to be on providing a detailed map of the current job market for the most relevant industries, and develop a forward-thinking perspective on job opportunities and skill demand trends.

Data Set

The dataset in question is made up of two primary CSV files which are entitled `job_skills.csv` and `linkedin_job_postings.csv`, comprised of 1,348,435

unique posings. Luckily, these entries are entirely non-null in all columns with the exception of only 19 entries throughout the whole dataset. The `job_skills.csv` file contains a comma-separated list of skills, as well as a URL foreign key for joining to the other table:

1. **job_link**: String (Nominal) - URL, Foreign Key
2. **job_skills**: String(Nominal) - comma-separated list of skills required for that job

The second table, `linkedin_job_postings.csv` holds the comprehensive details about each posting:

1. **job_link**: String (Nominal) - URL, Primary Key
2. **last_processed_time**: DateTime (Interval) - When posting was last processed
3. **got_summary**: Boolean (Binary) - Indicates if posting summary was successfully extracted
4. **got_ner**: Boolean (Binary) - Whether Named Entity Recognition was performed
5. **is_being_worked**: Boolean (Binary) - Status of posting
6. **job_title**: String (Nominal) - Title of posting
7. **company**: String (Nominal) - Company that is posting
8. **job_location**: String (Nominal) - Where the posting is located
9. **first_seen**: DateTime (Interval) - Date posting was made
10. **search_city**: String (Nominal) - City used to find posting

This massive dataset that is for the most part devoid of null values will set a strong foundation for impactful analysis and insights into current job market dynamics.

Evaluation Methods

I intend to compare my foundational findings to the existing job market analytics which were discovered in the 'EDA on LinkedIn jobs and skills', and 'Analysis of 1.3 Million Jobs on linked-in' Jupyter Notebooks. By utilizing their findings as a benchmark

I will assess the uniqueness and depth of my own insights. I also intend to quantify the coherence and distinctiveness of clusters within my analysis, adopting a threshold such as a silhouette score of at least 0.5 (a metric for measuring cluster quality) that ensures clusters are both internally consistent and distinctly separate from each other. This approach will ensure the robustness of the clustering methodology. Additionally, I plan on leveraging the Apriori algorithm, I will discover all of the frequent itemsets among skills and job titles across industry clusters, thus further informing this project's insights into the 2024 job market's structure and dynamics.

Part 3: Updated Evaluation Methods

As the project progresses, my approach to evaluating the effectiveness and accuracy of the analysis continues to be rigorous. Initially, I proposed to benchmark my findings against existing job market analytics derived from prominent Jupyter Notebooks such as 'EDA on LinkedIn jobs and skills' and 'Analysis of 1.3 Million Jobs on LinkedIn'. This comparison has been vital in understanding the uniqueness and depth of the insights my project brings to the table. By contrasting my results with these benchmarks, I have been able to highlight distinct patterns and trends not previously identified, including accurate and nuanced industry classification, affirming the value of this project's approach.

Furthermore, to ensure the reliability of my chosen clustering methodology, I will adopt the silhouette score as a metric for evaluating the quality of the clusters formed. I target a silhouette score of at least 0.5 to confirm that my clusters are internally consistent yet distinctly separate from each other.

Moving forward, these evaluation methods will continue to inform the refinement of this analysis. I aim to deepen the exploration of the data and extend the evaluation to include additional metrics and comparisons that will solidify the credibility and relevance of the project findings. This continuous measurement ensures not only the integrity of my research process but also increases the actionability of the insights generated.

Tools

I plan to utilize Python as well as its flagship data analysis libraries including Numpy, Pandas, Matplotlib, and Seaborn. I intend to utilize Git and GitHub for version control and regular project updates. For any formulas and techniques, I intend to use 'Data Mining Concepts and Techniques - Third Edition.'

Milestones

I plan to complete work on this project at regular intervals in accordance with class deadlines:

- **03/18-03/25** - Begin data collection and initiate data cleaning to remove sub-par postings.
- **03/26-04/02** - Conduct validation on subsets of postings and integrate postings & skills.
- **04/03-04/10** - Classify job titles and generate a structured format suitable for analysis.
- **04/11-04/18** - Conduct time-series analysis, identifying trends in regional industry booms and job title popularity.
- **04/19-04/26** - Employ clustering to group similar job titles and associated skills.
- **04/27-05/01** - Complete analysis and compile findings.

Part 3: Milestones Completed

- **03/18-03/25: Data Collection and Cleaning:** Successfully collected the "1.3 M LinkedIn Jobs & Skills (2024)" dataset and began thorough data cleaning to remove subpar postings, ensuring the data quality for subsequent analysis.
- **03/26-04/02: Data Validation and Integration:** Conducted detailed validation on selected subsets of postings to verify their integrity and integrated job postings with corresponding skills, setting up for structured analysis.
- **04/03-04/10: Job Title Classification:** Effectively classified job titles into broad industry sectors and generated a structured data format suitable for nuanced analysis,

facilitating easier identification of industry trends.

- **04/11-04/18: verification of Classification:** Validated industry classification facilitated through a comparison between most common industry skills and posting skills, removing mismatched classifications.

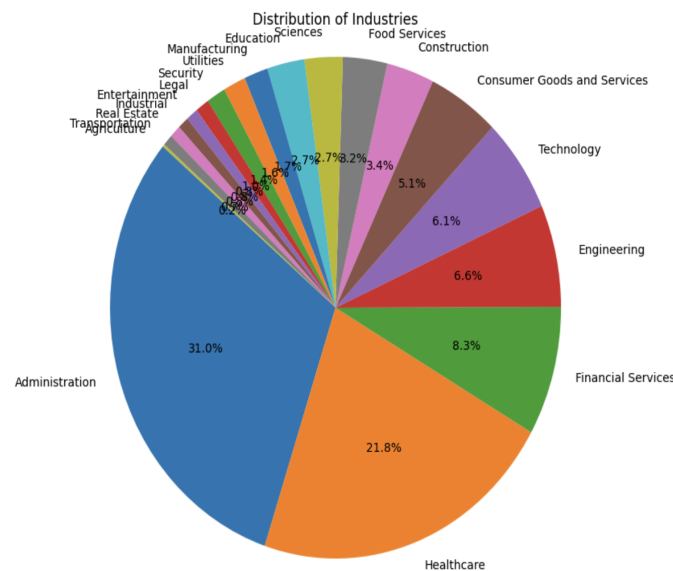
Part 3: Milestones Todo

- **04/19-04/26: Clustering Job Titles and Skills** and utilize k-means clustering to group similar job titles and associated skills within industry sectors, uncovering significant patterns in skill demand across different regions and job categories
- **04/27-05/01: Final Analysis and Report Compilation:** Complete the final stages of analysis, synthesizing all findings into a set of final insights. Compile the final report, incorporating all analyses, visualizations, and evaluations into the final report. Prepare a presentation to summarize the project's insights and contributions, aiming to highlight the practical applications of the findings for job seekers, educational institutions, and industry professionals.

Results So Far

The current analysis of the "1.3 M LinkedIn Jobs & Skills (2024)" dataset has already yielded many significant insights into the job market for 2024. The data, once cleaned and structured, provided a revealing insight into key trends and patterns that inform current and future job seekers, and educators. Two primary results stand out from the analytical processes applied so far:

Industry Distribution Analysis:



This pie chart visualizes the proportionate makeup of job postings across various industries within the dataset. This visual representation helps to quickly grasp which industries are dominating the job market in 2024, showing several clear leaders in terms of market opportunities. certain industries including Administration, Healthcare, and Finance show significant representation, indicating robust growth and demand for skills in these areas.

Top Skills by Industry:

This table lists the top 10 skills identified within each industry, derived from the skillset analysis. This table is crucial for understanding the specific skill demands of high-growth industries. For instance, in the Technology sector, skills like SQL, Python, and Data Analysis top the list, whereas in Healthcare, Nursing, Patient Care, and Documentation are most sought after. This alignment of job market needs with skill sets provides a direct benefit to job seekers and curriculum developers aiming to meet industry standards and demands.

These results not only validate the effectiveness of the methodologies employed (such as keyword mapping and skillset comparison) but also provide actionable insights. They illustrate a clear and detailed picture of the 2024 job market's structure and dynamics, which is essential for strategic career planning and educational alignments.

Industry	Top 10 Skills
Administration	Communication, Leadership, Teamwork, Customer service, Customer Service, Training, Problem Solving, Project Management, Communication skills, Sales
Agriculture	Customer service, Communication, Sales, Teamwork, Marketing, Communication skills, Leadership, Insurance, Customer Service, Problem solving
Construction	Communication, Project Management, Leadership, Scheduling, Teamwork, Construction Management, Problem Solving, Troubleshooting, Budgeting, Construction
Consumer Goods and Services	Customer service, Communication, Teamwork, Customer Service, Sales, Leadership, Merchandising, Communication skills, Problemsolving, Problem Solving
Education	Communication, Teaching, Teamwork, Collaboration, Leadership, Research, Communication skills, Education, Problem Solving, Curriculum Development
Engineering	Communication, Project Management, Python, Teamwork, Troubleshooting, Leadership, Electrical Engineering, AutoCAD, Engineering, Mechanical Engineering
Entertainment	Communication, Teamwork, Customer service, Collaboration, AutoCAD, Leadership, Project Management, Communication skills, Sales, Attention to detail
Financial Services	Communication, Sales, Customer service, Teamwork, Customer Service, Accounting, Communication skills, Leadership, High school diploma, Microsoft Office Suite
Food Services	Communication, Teamwork, Customer service, Leadership, Customer Service, Sanitation, Food preparation, Food safety, Food Safety, Training
Healthcare	Nursing, Communication, Patient Care, Teamwork, Patient care, Leadership, Registered Nurse, Documentation, BLS, Collaboration
Industrial	Communication, Leadership, Teamwork, Logistics, Problem Solving, Inventory Management, Warehouse Management, Customer Service, Customer service, Microsoft Office Suite
Legal	Litigation, Legal Research, Communication, Teamwork, Legal research, Legal Writing, Negotiation, Case Management, Attention to Detail, Communication Skills
Manufacturing	Communication, Leadership, Manufacturing, Teamwork, Problem Solving, Project Management, Safety, Troubleshooting, Quality Control, Scheduling
Real Estate	Communication, Customer service, Leasing, Property Management, Sales, Customer Service, Marketing, Teamwork, Microsoft Office Suite, Negotiation
Sciences	Communication, Teamwork, Nursing, Leadership, Research, Collaboration, Patient Care, Data Analysis, Communication skills, Data analysis
Security	Communication, Teamwork, Leadership, Project Management, Security, Python, Communication skills, Communication Skills, Troubleshooting, Bachelor's Degree
Technology	Communication, SQL, Python, Teamwork, Data Analysis, Project Management, Java, Problem Solving, Collaboration, Leadership
Transportation	Communication, Teamwork, Customer service, Leadership, Customer Service, Communication skills, Microsoft Office, Problem Solving, Project Management, Scheduling
Utilities	Communication, Project Management, Leadership, Teamwork, Engineering, Problem Solving, Project management, Troubleshooting, Communication skills, Communication Skills

Moving forward, these initial findings will be further explored through deeper analytical methods, such as excluding common soft skills, and potentially uncovering more nuanced interactions between job titles, skills, and industry trends. This continued analysis will refine these results, adding new layers of interpretation and recommendations based on emerging patterns.

Part 4: Abstract

At the onset of this project, I sought to paint a clearer picture of the 2024 job market, and provide tools to help job seekers and educational institutions better understand current skill demands and industry hotspots. The result of this effort is a set of modular tools that offer invaluable insights by Industry, and pave the way for answering questions such as:

- In what regions and cities are jobs in my industry booming?
- What are the most popular skill sets required among postings in my industry?
- What skills in my industry are currently in high demand?

In brief, The results of this project are a set of tools to map your career in any of 18 industries which each posting was systematically classified as by systematically refining an industry-keyword map. Within the Technology Industry, the most in-demand

skills are SQL, Python, and Java, and the most frequent skill sets are (AWS, Kubernetes, and Docker), as well as (SQL, and Python). Some of the top Cities for Tech in the nation are San Francisco and New York, and some of the most popular cities in the state of California are San Francisco and San Jose. This level of analysis is available for 18 different industries and has such has the potential to help a wide range of job seekers and educational institutions to gain valuable and actionable insights.

Part 4: Introduction

As a job seeker, or educational institution, it can be difficult to stay in touch with the ever-changing demands of the world's rapidly evolving industries. People need to remain in touch with the worlds fast paced job market, and this project provides a set of tools that can answer some very powerful questions. One's entire life is dependent on where they live from the friends that they will make to the weather outside. As such, it is important to make careful considerations about answering the question: 'Where do I want to live?' It is important that the answer to this question supports one's career goals, and as such, having a nuanced understanding of where your industry is thriving is vital.

On top of making sure that one is living where their chosen industry is thriving, it is doubly important that one is staying on top of their industries rapidly changing list of in-demand skills. Without asking questions like, 'What are the most popular skills in demand in my industry?', One can set themselves up for failure by learning dated or simply out-of-demand skills. Learning in-demand skills will always remain a cornerstone of maintaining relevance as a job seeker.

In addition to making sure that one is learning only the most relevant in-demand skills, one should also be asking themselves, 'What skills are often required in tandem?' Having an in-demand skill is great, but a career is an interweaving of in-demand skills that correlate highly and work well with each other. As such it is important that job seekers are not only learning in-demand skills, but in-demand skills that are more valuable when learned together.

Part 4: Data Set

In addition to the "1.3 M LinkedIn Jobs & Skills (2024)" dataset which was acquired on Kaggle, Additional data was utilized in order to reveal geospatial insights. The `uscities.csv` dataset courtesy of `simplemaps.com`, contains latitude and longitude lines for the majority of the cities in the US and was utilized to geocode job postings based on the city the posting was from. In addition to `uscities.csv`, I also utilized a shapefile of the continental United States, which was invaluable for plotting the geocached postings based on the latitude and longitude lines of each posting.

Part 4: Main Techniques Applied

Data Cleaning and Transformation

At the heart of this data-driven project lies the imperative process of data cleaning and transformation, which sets the stage for all analysis. In this project, I meticulously curated the datasets to ensure the accuracy and relevance of the insights generated. This phase involved several steps designed to refine and optimize the data for analysis. Initially, I removed all rows with null values from the primary datasets, `job_skills_df`, and `linkedin_job_postings_df`, to ensure the completeness and reliability of the data. Recognizing the importance of coherence between datasets, I next ensured that all job links in `linkedin_job_postings_df` existed within `job_skills_df`, thus aligning both datasets by their fundamental identifiers. To further streamline the data, I identified and removed job links that did not match between the two datasets. This deletion of unmatched job links prevented any discrepancies in the later stages of analysis. Additionally, I removed less relevant columns such as `'last_processed_time'` and `'search_position'`, which were irrelevant for this analysis, thereby focusing on the most impactful data attributes.

Another important technique was the transformation of job skills data into a clean, deduplicated list of skills for each job posting. By converting the skills data into a standardized format—lowercased and stripped of extraneous whitespace—we ensured uniformity and facilitated easier aggregation and comparison. This transformation was accomplished using a custom function `clean_skills`, which processed

each string of skills into a concise list of unique skills, thus enhancing the clarity and utility of our skills data. Finally, the datasets were merged based on the `job_link` column, aligning all related data into a unified frame. This merge not only consolidated our data landscape but also set the stage for robust, in-depth analysis across various dimensions of the job market.

Industry Classification

The classification stage of this project was essential in correlating job postings to their respective industries, which is a foundational step toward uncovering industry-specific trends. This process was defined by the development and refinement of a comprehensive keyword-industry map. Starting with an initial set of industry-specific keywords, I expanded the map iteratively, adding keywords manually to enhance the coverage until over a million postings were successfully categorized into one or more of the eighteen key industries we identified.

Using a `'categorize_job'` function, I mapped each job title and company name to potential industries by scanning and matching against the keyword map. This approach ensured that each posting was evaluated for relevance to predefined industry categories, using a combination of job title and company name.

Classification alone does not guarantee the accuracy of the results and validation is equally crucial. To this end, I implemented a validation step to ensure that the skills listed in job postings matched the most common skills identified for their classified industries. The `'get_top_skills'` function was important here, identifying the top skills for each industry by analyzing the frequency of skills across postings within each industry. Then, using the `'validate_job_posting'` function, each job posting was assessed to determine whether it contained any of the top skills identified for its respective industry.

The approach of classification followed by validation was effective in filtering out inaccuracies. Out of the initial set of categorized postings, 1,110,590 were validated as valid fits within their industries, while 448,937 were identified as mismatches and subsequently removed from consideration. These statistics highlight the robustness of the methodology, as well as its ability to refine the categorization based

on concrete skill data. By ensuring that each posting not only fits an industry by keywords but also aligns with the skill demands of that industry, I have significantly enhanced the reliability of the analytical outputs.

Frequent Skill Sets with Apriori

To dive deeper into the collaborative nature of skills within industries, I utilized the Apriori algorithm, a popular method within the field of data mining for extracting frequent item sets. This approach is particularly beneficial for the purpose of identifying common combinations of skills that frequently occur together within job postings, providing a granular view of the skill synergies that are most valued across different industries.

This process began with a more focused data sampling strategy restricted by the computing power and memory constraint of my JupyterHub environment, where I reduced the dataset to manageable proportions while maintaining a representative sample. This sampling was critical to ensure efficiency and effectiveness in identifying significant patterns without the overhead of processing the entire dataset. Using the `'sample_data'` function, I randomly selected subsets of data, ensuring diversity and randomness in the samples to avoid biased insights.

Following data sampling, the next step was to transform the skills data into a format suitable for mining frequent item sets. I utilized `TransactionEncoder` to convert lists of skills into a sparse matrix format, where each row represents a job posting and each column a skill, encoded as present or absent. This transformation, executed through the `encode_data` function, was instrumental as it prepared the data for the subsequent application of the Apriori algorithm.

Using the `'apply_apriori'` function, I applied the Apriori algorithm to the encoded data, setting a minimum support threshold that determines the relevance of an item set based on its frequency of occurrence. The results revealed combinations of skills, indicating the core skills that are often desired together in the industry. For example, in the Technology industry, combinations such as (Python, and SQL) and more complex sets like (Python, Java,

and SQL) were identified as frequent, highlighting the demand for these skills in synergy.

The final step was sorting these item sets by the number of skills they contain, prioritizing those with higher cardinality, thus suggesting a strong and more significant association between the skills within the set. This sorted list not only highlights the most in-demand skill combinations but also provides insights into the evolving complexity of skill requirements in the 2024 job market.

Geospatial Industry Boom Mapping

The visualization of geospatial data is important in understanding regional dynamics and trends within various industries. In this project, I focused on mapping the distribution of job postings across the United States at both national and state levels, employing geospatial techniques to reveal the hotspots of industry activity. This approach not only highlights regions with high job densities but also aids job seekers and educational institutions in identifying areas of booming industry-specific employment.

Utilizing the `'plot_industry_job_postings_US'` function, I was able to filter and map the data specific to each industry, such as Technology and Healthcare, overlaying job posting counts onto a US map shapefile. By merging the job data with coordinates from the `'us_cities_df'` dataset, I upgraded the analysis with precise location mapping. Each city's job posting count was then represented on the map with proportional dot sizes and annotations for the top cities, enhancing the visual representation of the data and making it easier to identify areas with the highest demand.

For example, in the Technology industry, the analysis highlighted New York, San Francisco, and Seattle as the top cities with the highest number of job postings, indicating these cities as important hubs of tech employment. Similarly, in the Healthcare industry, Los Angeles, New York, and Boston emerged as the top cities, underscoring their significant roles in healthcare employment.

Further refining the geographic analysis, I utilized the `'plot_industry_job_postings_State'` function to narrow in on state-specific data. This allowed for the mapping of job postings within individual states, such as California for Technology and Colorado for

Healthcare, thus providing a more localized view of job distribution. In California, San Francisco, Los Angeles, and San Diego were identified as the best cities for technology jobs, whereas in Colorado, Denver, Aurora, and Colorado Springs were the best for healthcare jobs.

Part 4: Key Results

This analysis, conducted on the LinkedIn job market, and for practical example, the Technology industry, has unearthed many valuable insights that hold considerable real-world applications for job seekers, particularly those aspiring to roles within software engineering and related fields. Through this comprehensive examination of job posting data, I have discovered several key factors that are essential for career advancement in this fast-paced industry.

Firstly, this study highlighted the most in-demand skills, with SQL, Python, and Java emerging as pivotal competencies sought by employers. The significance of these skills underscores the need for prospective job candidates to attain proficiency in these areas to enhance their employability and competitive edge in the job market. In addition, the identification of important skill sets, such as combinations of (SQL, and Python) and (AWS, Kubernetes, and Docker), provides job seekers with clear guidance on the capabilities that are currently valued highly across the Technology industry.

In addition to skill-specific insights, the geospatial analysis revealed the geographical distribution of technology job opportunities across the United States, indicating a very high concentration in the coastal regions, specifically in the East and West. California was a standout in this analysis and was subjected to a deeper state-level analysis, which confirmed that cities like San Francisco, San Diego, San Jose, Los Angeles, and Irvine are hotspots for technology roles. This geospatial insight is very valuable for job seekers, keeping them informed about where they might consider relocating or focusing their job search efforts to maximize their chances of securing employment.

The implications of these findings also extend from individual job seekers to educational institutions that play a crucial role in preparing future job seekers. By aligning their curriculum with the skills and

technologies that are in high demand, educational institutions can ensure that their offerings are relevant and that their graduates are well-prepared to meet the current and future needs of the Technology industry.

Part 4: Applications

Job Seekers

Job seekers, especially those entering or transitioning within the Technology industry, can utilize these insights to tailor their career development strategy effectively. Understanding the most in-demand skills and where they are needed can guide individuals in choosing specific certifications or training. For example, a junior software engineer might focus on mastering Python and SQL, and consider relocating to tech hubs such as San Francisco or San Jose to increase their employment prospects. Additionally, the knowledge of popular skill combinations, like AWS with Kubernetes and Docker, can help candidates align their learning paths with the synergies that employers most frequently require.

Educational Institutions

Educational institutions, from universities to other vocational training, can utilize these findings to keep their curriculum up to date and ensure that program offerings better align with industry demands. By integrating courses that focus on high-demand skills and technologies identified in this analysis, these institutions could increase the employability of their graduates. In addition, the data-driven understanding of geographic job distribution gives these institutions the ability to establish targeted remote learning opportunities in regions that have abundant job postings, thus making education more relevant and accessible.

Industry Recruiters

Recruiters and HR in the technology sector can apply these insights to refine their talent acquisition strategies. Having knowledge of the most sought-after skills and the locations of job postings can inform targeted recruitment campaigns, allowing companies to source the best talent more efficiently. Thus,

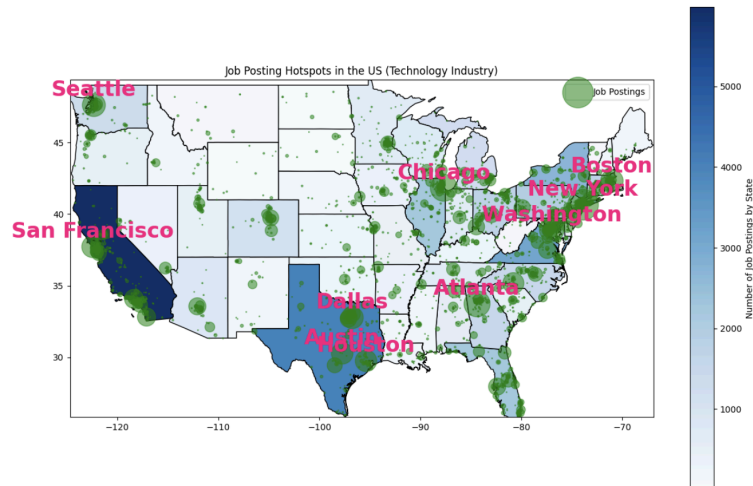
understanding the skill trends and regional job clusters can assist in workforce planning and development, thus ensuring that recruitment efforts are not only strategic but also aligned with the industry's evolving needs.

Policymakers

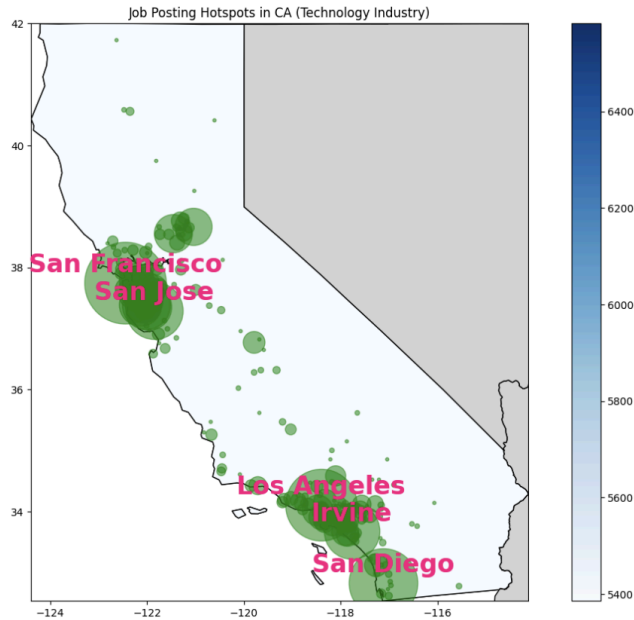
Policymakers and economic planners can leverage this data to promote economic development in their regions. Insights into the regions with high concentrations of tech jobs can guide potential infrastructure investments, such as the development of tech parks or enhanced internet connectivity, to support and attract technology companies. In addition, these findings can help decide educational incentives aimed at skills development in high-demand areas, thus aligning workforce skills with market needs and boosting local employment rates.

Part 4: Visualization

National Technology Hotspots



California Technology Hotspots



References

- [1] Baoxu Shi, Jaewon Yang, Feng Guo, and Qi He. 2020b. Saliency and market-aware skill extraction for job targeting. (May 2020). Retrieved March 16, 2024 from <https://arxiv.org/abs/2005.13094>
- [2] Galal Qassas. 2024. Analysis of 1.3 million jobs on linked-in. (February 2024). Retrieved March 16, 2024 from <https://www.kaggle.com/code/galalqassas/analysis-of-1-3-million-jobs-on-linked-in/notebook>
- [3] Varnika Bhardwaj. 2024. EDA on linkedin jobs and Skills. (February 2024). Retrieved March 16, 2024 from <https://www.kaggle.com/code/varnikabhardwaj/eda-on-linkedin-jobs-and-skills>
- [4] United States Cities Database. simplemaps. (n.d.-a). <https://simplemaps.com/data/us-cities>