Stats401 Final Report:
Decoding the American Dream: Visual Analytics of Job Market in the U.S.

Jingheng Huan, Zhiyun Lu, Yantao Mei

Introduction

This report utilizes advanced data visualization techniques to decode the U.S. job market, offering key insights for international talents pursuing the American Dream. Leveraging diverse methods like radar charts and bubble maps, we analyze varied datasets, including Glassdoor and LinkedIn. Our visualizations address employment opportunities, salary metrics, and industry trends, simplifying complex data for informed decision-making. This serves as a comprehensive guide for students, professionals, and industry analysts to navigate the U.S. job market effectively.

Dataset

In pursuit of a comprehensive analysis, we have sourced two principal datasets from Kaggle's extensive repository. The first, a Glassdoor dataset, provides a wealth of information on salary distributions, education level, and employer reviews. The second, a LinkedIn dataset offers an almost exhaustive snapshot of 15,000+ job postings across the United States within a 48-hour span. To augment these datasets, we employed Google Map API to acquire precise geographical coordinates of U.S. cities, thereby enriching the 'location' attribute in our analysis. Furthermore, we utilized the Yahoo API to obtain financial metrics related to companies listed in the S&P 500, adding depth to our corporate evaluations. To provide geographical context, especially for our bubble map visualization, we incorporated a U.S. county map sourced from geojson. These multi-source data streams synergize to create a robust foundation for our visual analytics, enabling an intricate yet coherent understanding of the U.S. job market.

Word Cloud

To delineate the landscape of job titles in the U.S. job market, we utilized the LinkedIn dataset from Kaggle for this specific visualization task. Our focus was narrow yet deep, concentrating solely on job titles and their corresponding frequencies. To ensure the relevance and specificity of our analysis, we filtered the dataset to include only "full-time" job positions. In the resulting word cloud, we aimed to identify the top 30 most frequently occurring job titles on LinkedIn.

In terms of user interaction, we incorporated a versatile filter allowing users to sort the word cloud based on four different metrics: "frequency," "maximum salary," "median salary," and "minimum salary." After selecting, a hoverboard feature is activated, displaying detailed attributes of the chosen job titles, thereby enhancing the interactive experience and informational value of the visualization.

Our analysis yielded intriguing findings. For instance, the job titles "Sales Director" and "Owner and Operator" emerged as the most frequent, while "Quantitative Developer" positions boasted the highest salary range. Notably, sectors like Engineering and FinTech, along with managerial

roles, were not only prevalent but also associated with competitive salaries. These insights illuminate the complexities of the job market, offering actionable intelligence for job seekers and industry analysts alike.

Radar Map

To gain a nuanced understanding of corporate performance within the U.S. job market, we implemented a radar map visualization. This effort relied on data obtained through the Yahoo API, focusing on companies listed in the S&P 500 index, complemented by information from the LinkedIn dataset. One of the remarkable aspects of the acquired data was its completeness, obviating the need for additional data cleaning or imputation.

The radar chart is designed to visualize multiple key financial metrics: volume, current price, total cash per share, book value, current ratio, and gross margin. This multi-metric approach offers a holistic view of a company's financial health, thereby aiding in comprehensive analysis.

The visualization incorporates a dynamic filter, enabling users to select and compare the financial performance of multiple companies simultaneously. Upon selection, the hues and scales are dynamically adjusted to ensure optimal readability and interpretability. Each financial attribute is also annotated with tooltips, offering users immediate context and elaboration.

Our analysis led to several notable observations. For instance, NVR Inc outperformed other companies in nearly all financial metrics, while Bank of America took the lead in trading volume. These insights not only serve as a valuable resource for potential investors but also provide job seekers with an understanding of the stability and growth potential of prospective employers.

Bubble Map

To geographically contextualize job opportunities within the United States, we deployed a bubble map visualization using the LinkedIn dataset. During the preprocessing phase, our focus was solely on the 'location' attribute, leading us to filter out other features for this specific visualization. Locations ambiguously labeled as "US" or "United States" were excluded for the sake of accuracy. Additionally, we consolidated similar locations, such as combining "NYC" with "NYC Metropolitan area" and "Boston" with "Greater Boston," standardizing the data in a "city, state" format like "Durham, NC."

To further enhance geographical specificity, we incorporated a U.S. county map sourced from geojson. Geographical coordinates for each city were obtained using the Google Map API, allowing us to accurately position the bubbles on the map. The size of each bubble is proportional to the number of job opportunities available in that city.

Interactivity was a key focus in this visualization. We integrated a filter enabling users to categorize cities based on the number of job opportunities: "All," "0-50 jobs," "50-150 jobs," and "150+ jobs." Furthermore, a hoverboard feature provides detailed information about each city when hovered over by the cursor.

Our bubble map reveals a concentration of job opportunities in major U.S. cities like New York, Chicago, Houston, Los Angeles, and San Francisco, thereby offering valuable insights for job seekers and policymakers alike.

Ridge Line Plot

In our quest to explore the factors affecting salary distribution in the U.S. job market, we employed a ridge line plot, utilizing the Glassdoor dataset. The crux of this visualization is to understand the relationship between salary and education levels. To this end, we filtered out irrelevant attributes, focusing solely on education levels: high school, college, masters, and PhD.

Upon generating the ridge line plot, our findings defied conventional wisdom. The salary distributions across different education levels were not as distinct as anticipated. Contributing factors include the academic status of those labeled 'PhD' and the longer work tenures of individuals with only high school or college education.

A unique interactive feature of this plot is its focus on Kernel Density Estimation (KDE), which underlies the visualization. Users can interact with the plot by adjusting two KDE parameters using sliders: the kernel value and the smoothing parameter. This allows for a more tailored analytical experience.

It's worth noting that the data source itself has limitations. Though Glassdoor is generally reliable, the self-reported nature of data like age, salary, and education level can introduce variability, affecting the robustness of our analysis.

Tree Map

For our tree map visualization, we harnessed multiple data sources from the LinkedIn dataset and employed Tableau for its adeptness in data integration, mapping the unique IDs in companies.csv to their corresponding industry and employee count. Where employee_counts.csv featured multiple entries for a single company, we opted to characterize the employee count by the highest reported figure, accounting for potential shutdowns or downsizing within the year.

The visualization was executed using D3.js, with guidance drawn from existing tree map tutorials on the D3 official site. The tree map itself is interactive, allowing users to select specific industries. This triggers an indexing function that redraws the graph to focus on the selected industry's subtree. Each cell is annotated with the company name, provided the cell's dimensions permit it. Tooltips furnish additional details, such as the industry and employee count for each company.

While the visualization successfully meets its objective of identifying companies and industries by employee size, it is not without limitations. Notably, the graph experiences some degree of overplotting, particularly when zoomed into specific industries. This is an acknowledged

shortcoming, yet the overall visualization offers a valuable lens through which to explore the size and scope of various industries and companies.

Roles and responsibilities:
Yantao: word cloud, radar map, ridge line, dashboard
Jingheng: bubble map, ridge line, report, poster
Zhiyun: tree map, report, poster

Conclusion
This report, titled "Decoding the American Dream: Visual Analytics of Job Market in the U.S.," serves as a multifaceted lens through which to view the complexities of the U.S. job market. Utilizing diverse datasets and a variety of visualization techniques, we offer a robust analysis, exploring key variables such as job titles, corporate financial metrics, geographical job distribution, and salary trends against educational levels. Our findings, while insightful, are not without limitations, primarily owing to data veracity and visualization constraints. Nevertheless, the visual analytics deployed here stand as powerful tools for international students, job seekers, and industry analysts. As we lift the veil on the dynamics of the U.S. job market, we contribute to the ongoing quest to better understand and navigate the pursuit of the American Dream.

References
https://www.kaggle.com/datasets/nilimajauhari/glassdoor-analyze-gender-pay-gap/data
https://www.kaggle.com/datasets/arshkon/linkedin-job-postings
https://github.com/kjhealy/us-county/blob/master/data/geojson/gz_2010_us_050_00_500k.json
https://developer.yahoo.com/api/
https://developers.google.com/maps