ClAIrvoyant: Establishing an Anchor Client

Alex Hill, Avery Wang, Vaibhav Jha

MADS ‘26

6 October 2025

Abstract

We were tasked with selecting an anchor client to launch and promote ClAIrvoyant, an AI recruitment tool designed to improve the hiring process and optimize company resources. The perfect client would be a market leader in a thriving sector that can benefit from optimizing their talent acquisition process. To help choose this client, we analyzed over 124,000 LinkedIn job postings in 2023 and 2024 and historical stock data from Yahoo Finance, using techniques such as Principal Component Analysis (PCA), clustering, outlier analysis, and time series. We found that Information Technology, Healthcare, Staffing and Recruiting, Finance, and Software Development are the healthiest sectors in the United States. Further data analysis and market research will give us the information we need to select a final anchor client.

Introduction

The current hiring process in industry has much room to improve. Companies must invest significant capital in applicant tracking systems (ATS) to filter through applications. Meanwhile, candidates must apply to large numbers of jobs with slim hopes of being considered, let alone hired. It is difficult for companies to assign recruiters to filter through many applications manually. ATS tools allow companies to quickly screen applications by skills, years of experience, and more. However, they have their flaws. Human labor is still required in later stages of hiring, siphoning valuable resources from the company that could be better spent on current projects.

Companies can avoid this problem and save valuable resources using artificial intelligence (AI). Good AIdeas created ClAIrvoyant, an AI tool that parses resumes and identifies candidates with the potential to stay and develop in the company long-term. To help launch this product, Good AIdeas needs an anchor client to establish their legitimacy and unlock more client opportunities. To identify an anchor client, we developed three core research questions:

1. Which sectors are "healthy?"
2. How can we identify market leaders?
3. Which companies have room to improve in hiring/retaining talent?

We seek a company with a highly successful track record in their respective industry. An ideal candidate would be a market leader with room to optimize their talent acquisition process.

Data

Our dataset contains information on more than 124,000 job postings from 2023 through 2024. The main CSV file contains each job title, company name, description, salary, number of views, number of applications, and more. A separate file contains information about the companies' industries, specialities, and employee counts at various time periods. Other files contain information on job benefits and the skills the job posting demands.

There are many issues with missing company names. Since we would like to identify companies who need help to hire candidates, we decided to exclude these job postings.

Our main variables of interest are:

- *Company name*: categorical nominal, there are nearly 25,000 different companies and 1719 postings do not have a company listed
- *Industry name*: categorical nominal, there are 389 different industries
- *Employee count*: quantitative continuous, the largest company has nearly 750,000 employees, but more than 75% of companies have less than 1,000 employees and the mean is about 2,500 employees
- *Company size*: discrete ordinal, a number from 1 to 7, with 1 being the smallest and 7 being the largest
- *LinkedIn follower count*: quantitative continuous, the maximum is over 3 million and the mean is about 80,000
- *Number of applicants*: quantitative continuous, the maximum is 967 and the mean is 10.59. There are over 100,000 missing values. This poses two major problems:
  - Many LinkedIn job postings that we have seen have over 100 applicants. This is clearly not reflected in the data.
  - 80% of the data is missing.
- *Number of views*: quantitative continuous, the maximum is 9975 and the mean is 14.62. There are 6,400 missing values, and the log of views is highly correlated with the log of applicants ($r = 0.845$), as shown in Figure 1.
- Other notable variables include benefits, salary, and company specialities. We used one-hot encoding for each job benefit and company speciality.

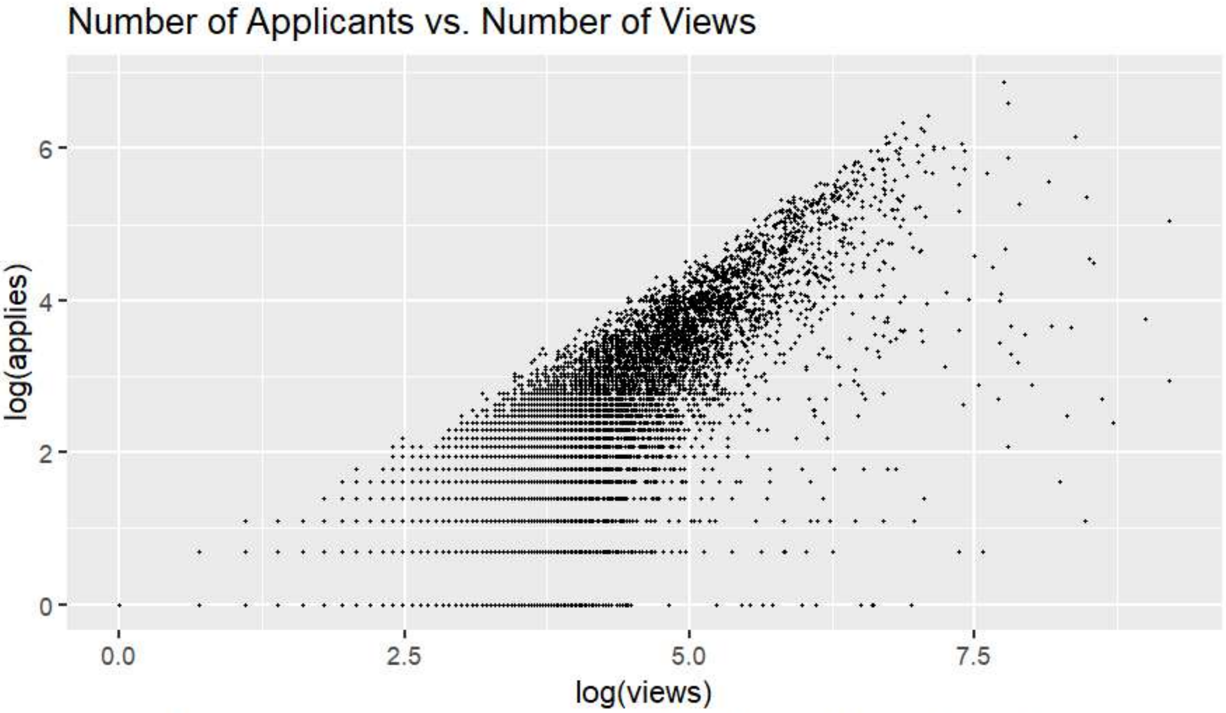

Figure 1: Number of applicants against the number of views for each job posting, plotted on a log scale.

We have a script to access historical stock data of any company, but because we have no way to find ticker symbols and we cannot check whether a company is publicly traded, we need to manually choose the companies for which to find stock data.

Methods

To fix our issue of missing values in the number of applicants, we fit a simple linear regression model of log of applicants on log of views and inputted the predictions where there was no true number of applicants. We chose not to use any salaries or benefits because there are too many missing values and benefit types. After this, we compiled the name, most recent employee count, number of LinkedIn followers, number of job postings, and number of applications (by posting and total) of every company.

We conducted an analysis of industries to evaluate sectors with high demand by finding the industries with the most job postings and most companies. These exploratory plots helped us understand the scope of our data; industries in the top ranks of these plots had a significant LinkedIn presence.

We continued our analysis of sector performance by performing principal component analysis. We felt that company size, employee count, follower count, posting views, and number of applicants strongly corresponded with an industry's health. To visualize potential associations, we created a biplot between the first two principal components, grouping by industry. Moreover, we could understand which variables had strong relationships with the top two principal components. Metrics with strong relationships with the principal components were deemed most relevant in assessing industry health.

Results

We noticed that the largest prediction outputted by the linear regression model was about 72 applicants. The vast majority of predictions were fewer than 10 applicants, indicating that this model did not affect our analysis substantially, since we focused on postings with more than 100 applicants.

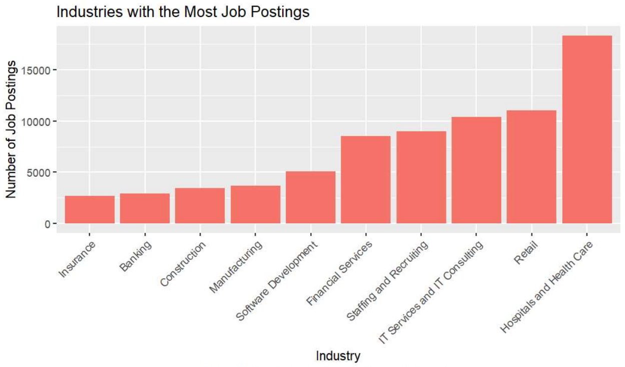

Figure 2: The ten industries with the most job postings.

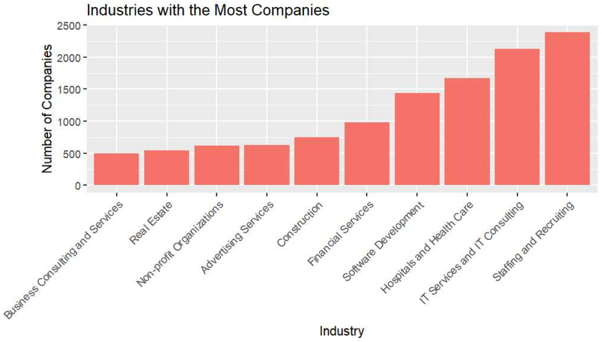

Figure 3: The ten industries with the most distinct companies in the data.

Figures 2 and 3 display industries with the most companies and job postings. Five industries fall in the top ten sectors for those metrics: Staffing and Recruiting, IT Services and IT Consulting, Hospitals and Health Care, Software Development, Financial Services.

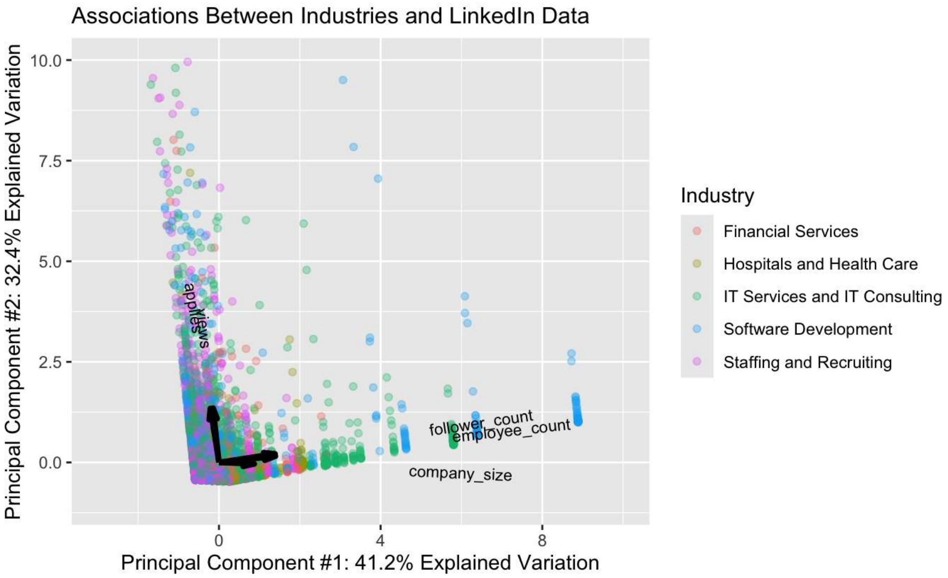

Figure 4: A principal component analysis modeling associations betweeen industries and LinkedIn data.

In **Figure 4**, we conducted Principal Component Analysis on five variables: job posting views, applications, company follower count, employee count, and company size. Next, we plotted the top two principal components, grouping companies by industry. The corresponding biplot indicates the statistics that measure the company itself. IT and Software Development tend to have larger follower counts, employee counts, and company size. On the other hand, the Finance, Healthcare, and Recruiting sectors have a positive relationship with views and applications.

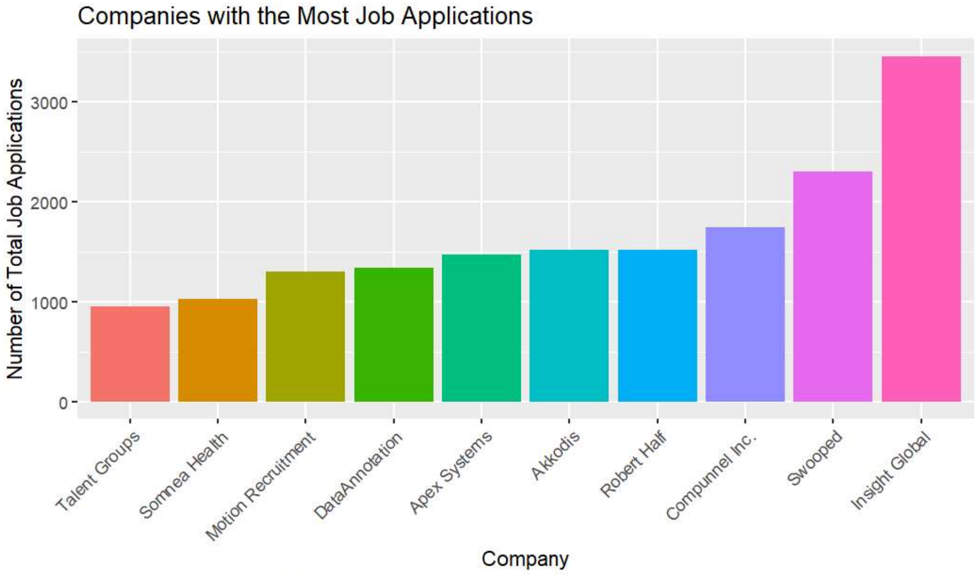

Figure 5: The ten companies with the most applications across all job postings in the data.

Figure 5 lists the companies with the most job applications. Insight Global, a Staffing and Recruiting firm with clients in IT, Healthcare, Finance, and other industries, has the most postings.

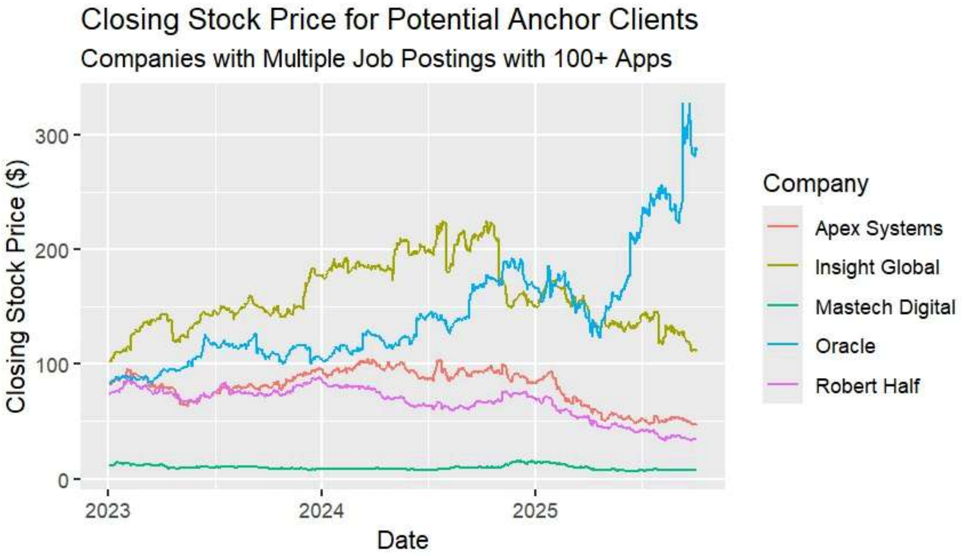

Figure 6: Time Series Plot for Stock Data of Potential Anchor Client Companies from 2023 to 2025.

Figure 6 displays stock data for Insight Global, Robert Half, Oracle, Apex Systems, and Mastech Digital, five companies with multiple job postings with at least 100 applications. Mastech Digital's stock price is much lower than the other four companies, while Oracle has the highest stock price. We plan to use stock data to identify more potential anchor clients.

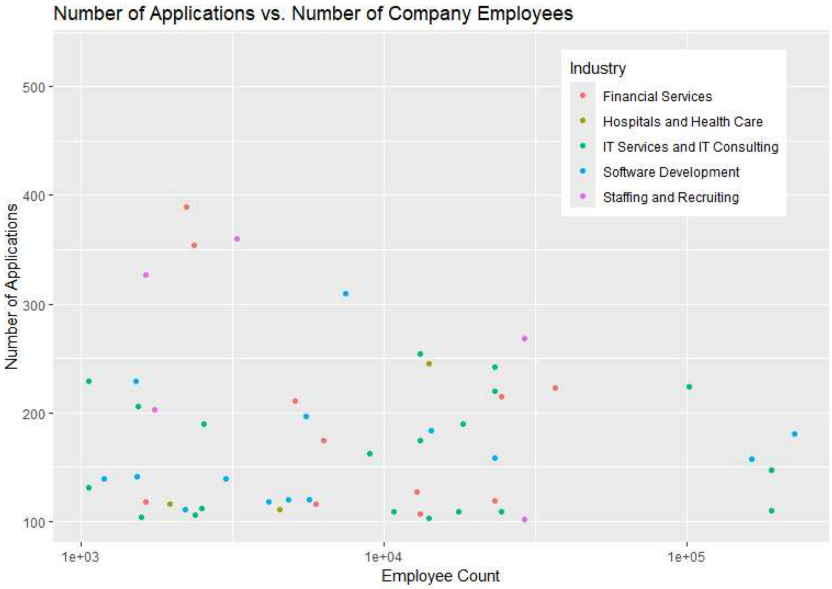

Figure 7: Scatter plot of most recent employee count (log scale, minimum 1000) against the number of applicants to individual postings (minimum 100).

Figure 7 plots the number of applicants to a job posting against the number of employees of the company, filtered by 100 applicants, 1000 employees, and the five industries of interest. This plot, along with **Figure 8**, the pair plot of the summary table, can help us identify clients that are market leaders with high demand for talent intelligence. For example, there are a few companies in Figure 7 with a large number of applicants in the Financial Services and Staffing and Recruiting industries. There are also a few companies with a large number of employees in the IT and Software Development industries. These are all worth further research, as companies with significant human capital needs have the most to gain from ClAIrvoyant.

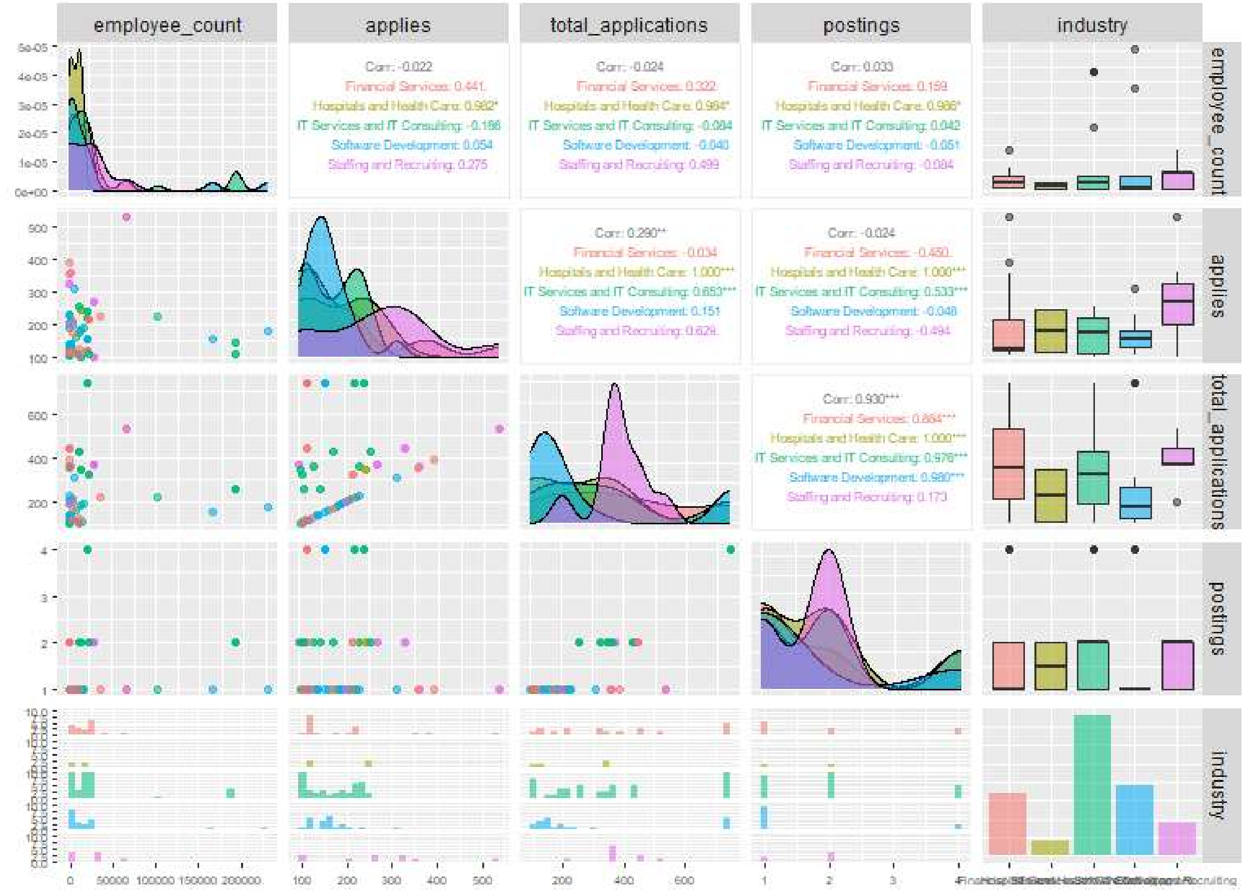

Figure 8: Pair plot of summary table.

Discussion

We identified Staffing and Recruiting, IT Services and IT Consulting, Hospitals and Health Care, Software Development, and Financial Services as the five healthiest sectors. Figures 2 and 3 show that these sectors overlap in having the most companies and the most job postings on LinkedIn.

At the company level, there are multiple options for choosing an anchor client. Insight Global is one option due to their high stock price and application volume (Figure 6). We will look into the other companies with well-performing stocks and outliers in variables such as job applications or number of employees. We will use market research of specific companies and top sectors to identify companies that have room to improve in talent acquisition and select an ideal anchor client for Good AIdeas and ClAIrvoyant.

There are various limitations to this analysis. First, the LinkedIn data excludes job postings and engagement on other platforms such as Indeed, Handshake, and the company websites. As a result, our dataset dramatically underestimates the number of applicants and views. In particular, the largest number of job postings with more than 100 applicants for a single company was only four, which is likely inaccurate.

Second, there are numerous missing values in the dataset, making it difficult to analyze. Our regression model to predict the number of applications has no way of capturing outliers, which is precisely what we are interested in. With more complete data, we can investigate more complex relationships to identify better anchor clients.

References

Arsh Koneru. (2024). LinkedIn Job Postings (2023 - 2024) [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/9200871.