Scraping Glassdoor Job Listings

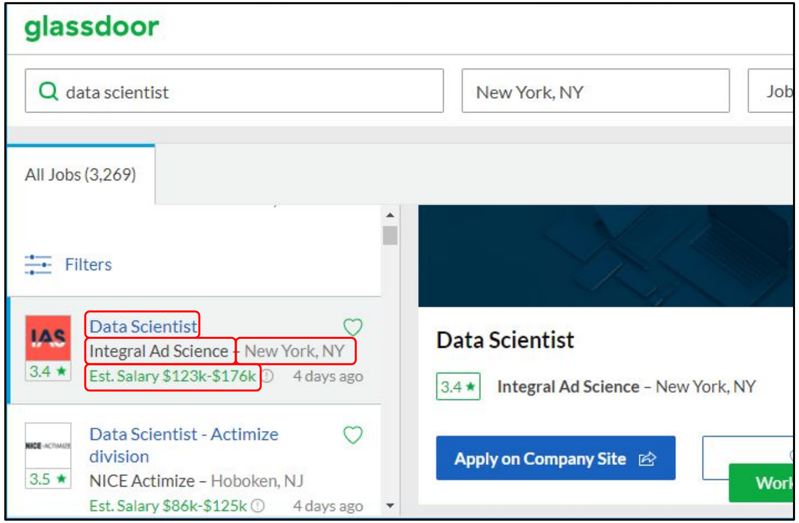
•••

Who wants to find a job?

John Merrick

The Plan: What to Scrape?

- Glassdoor job posting data using Scrapy
- Three-tier scraping algorithm:
  - Nine major metro areas
    - NYC, LA, SF, Santa Clara, Boston, Chicago, Seattle, Denver, Atlanta
  - Thirty Job Postings per search results page
  - 15 to 35 search results pages per metro area
  - Job post details for each listing
  - Eight data points:
    - Company, title, location, est. salary, rating, CEO approval, recommend friend, and job description



data scientist

New York, NY

All Jobs (3,269)

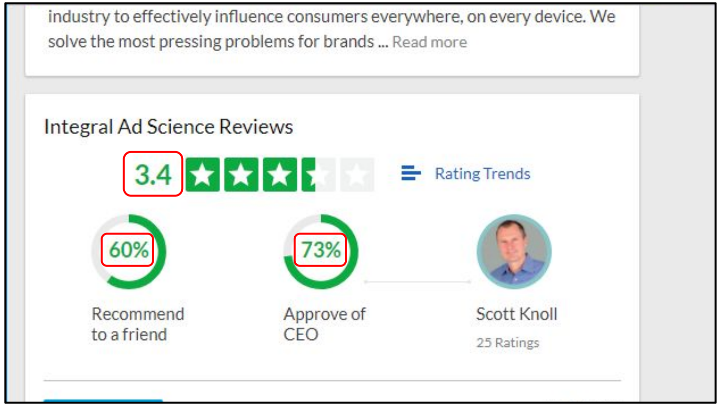
Data Scientist
Integral Ad Science – New York, NY
3.4 ★
Est. Salary $123k-$176k
4 days ago

Data Scientist - Actimize division
NICE Actimize – Hoboken, NJ
3.5 ★
Est. Salary $86k-$125k
4 days ago

Data Scientist
3.4 ★ Integral Ad Science – New York, NY

The Plan: What to Scrape?

- Returned ~7,500 job postings
- But this included ~2,500 clinical research scientist jobs



industry to effectively influence consumers everywhere, on every device. We solve the most pressing problems for brands ... Read more

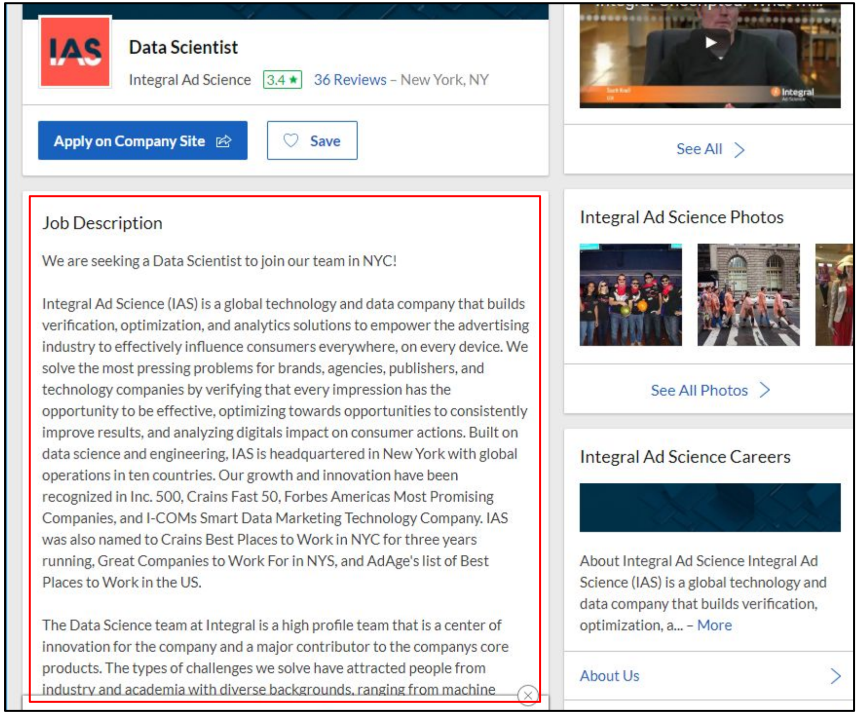
Integral Ad Science Reviews

3.4 ★★★★☆ Rating Trends

60% Recommend to a friend

73% Approve of CEO

Scott Knoll
25 Ratings



Data Scientist

Integral Ad Science 3.4 ★ 36 Reviews – New York, NY

Job Description

We are seeking a Data Scientist to join our team in NYC!

Integral Ad Science (IAS) is a global technology and data company that builds verification, optimization, and analytics solutions to empower the advertising industry to effectively influence consumers everywhere, on every device. We solve the most pressing problems for brands, agencies, publishers, and technology companies by verifying that every impression has the opportunity to be effective, optimizing towards opportunities to consistently improve results, and analyzing digitals impact on consumer actions. Built on data science and engineering, IAS is headquartered in New York with global operations in ten countries. Our growth and innovation have been recognized in Inc. 500, Crains Fast 50, Forbes Americas Most Promising Companies, and I-COMs Smart Data Marketing Technology Company. IAS was also named to Crains Best Places to Work in NYC for three years running, Great Companies to Work For in NYS, and AdAge's list of Best Places to Work in the US.

The Data Science team at Integral is a high profile team that is a center of innovation for the company and a major contributor to the companys core products. The types of challenges we solve have attracted people from industry and academia with diverse backgrounds, ranging from machine

Integral Ad Science Photos

Integral Ad Science Careers

About Integral Ad Science Integral Ad Science (IAS) is a global technology and data company that builds verification, optimization, a... – More

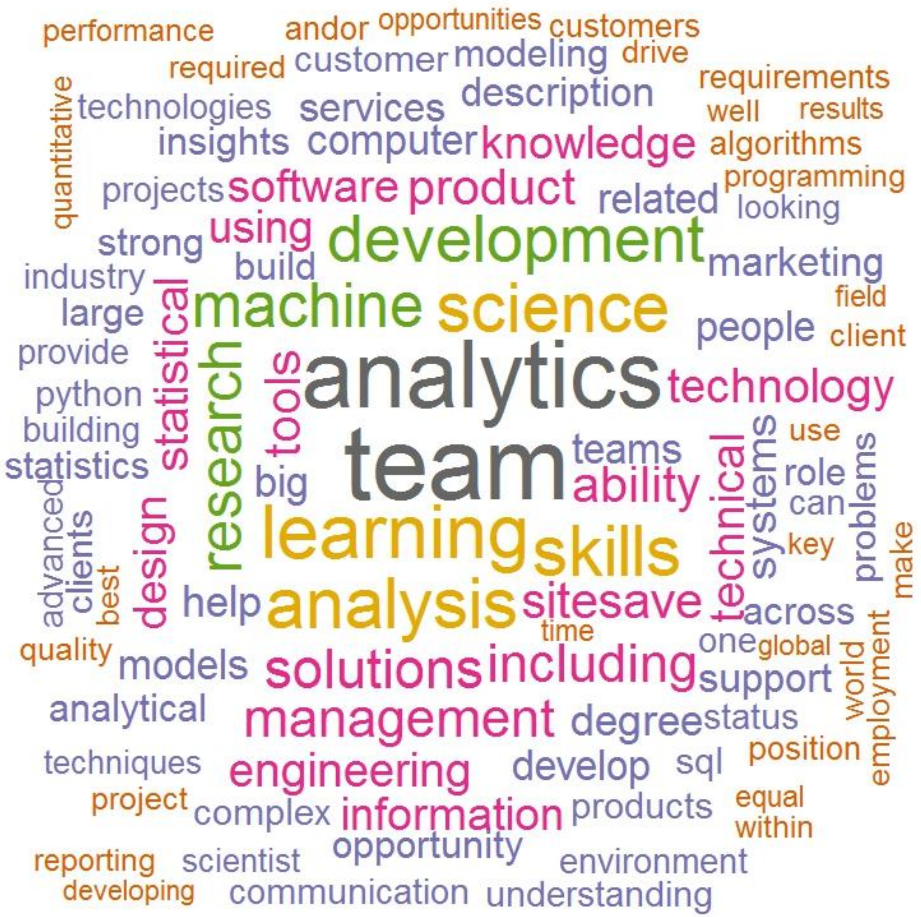
The Plan: Cleaning the data

- Extracted high and low est. salary to calculate midpoint and range
- Create T/F columns on keywords to eliminate scientific/clinical research postings
  - Biolog, pharm, clinic, immun, chemistry, oncology, biochemistry, neuro, medical, disease, physician, surgeon, nurse, hospital, cancer, vaccine, protein, specimen
  - If a posting had more than five of these indicators, I considered the job not relevant to the analysis and dropped the row
- Create T/F columns to determine if a job requires specific skills
  - Python, scikit, R, SQL, matlab, SAS, tableau, machine learning, natural language processing, hadoop, spark, java, mongo, hive, linux, statistics, visualization, PhD, and masters
  - If a posting had fewer than two of these indicators, I considered the job not relevant to the analysis and dropped the row

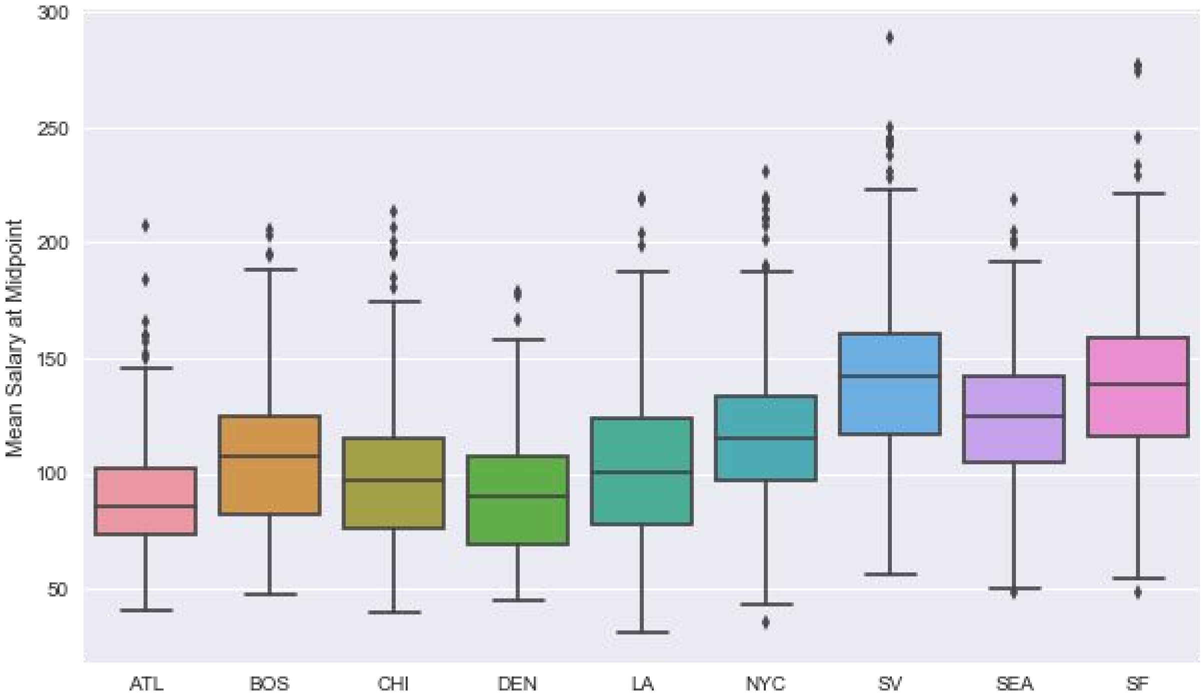
Analysis

Term Frequency

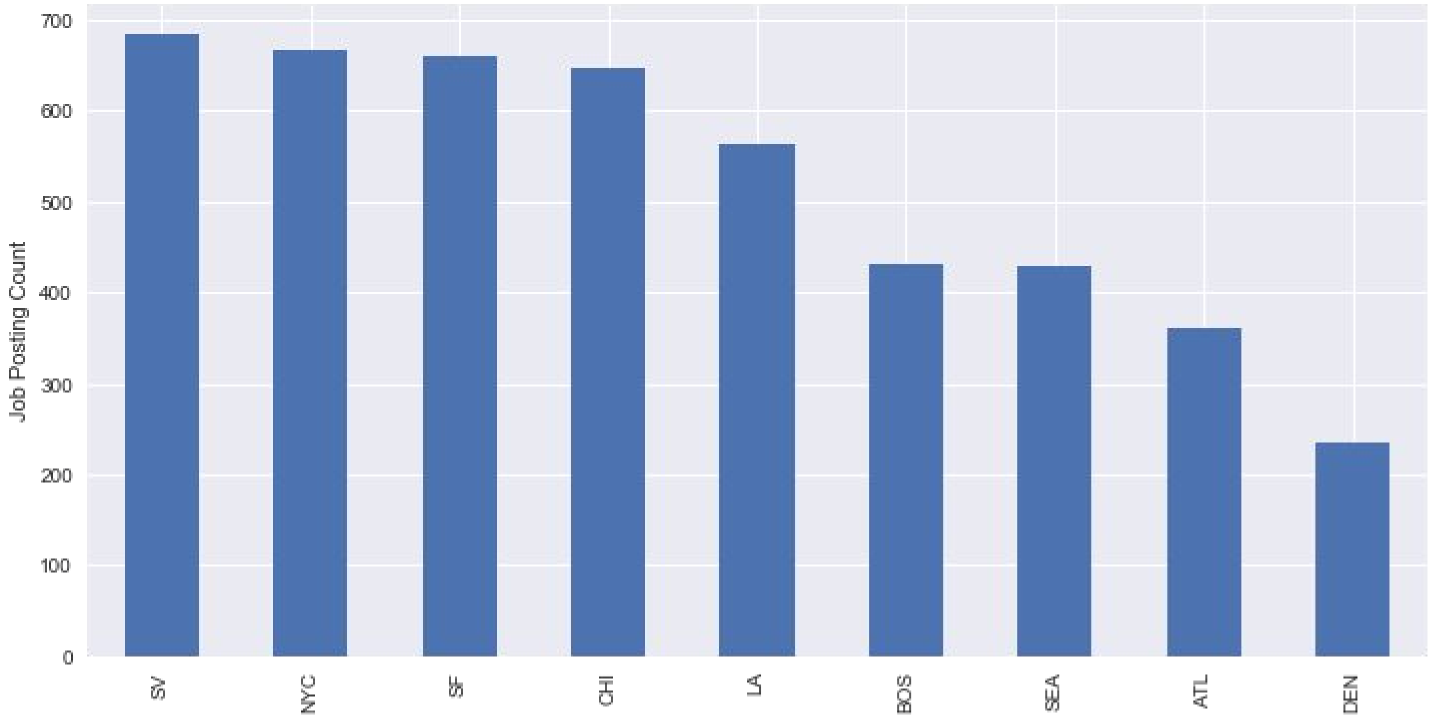
- Common terms across all job descriptions



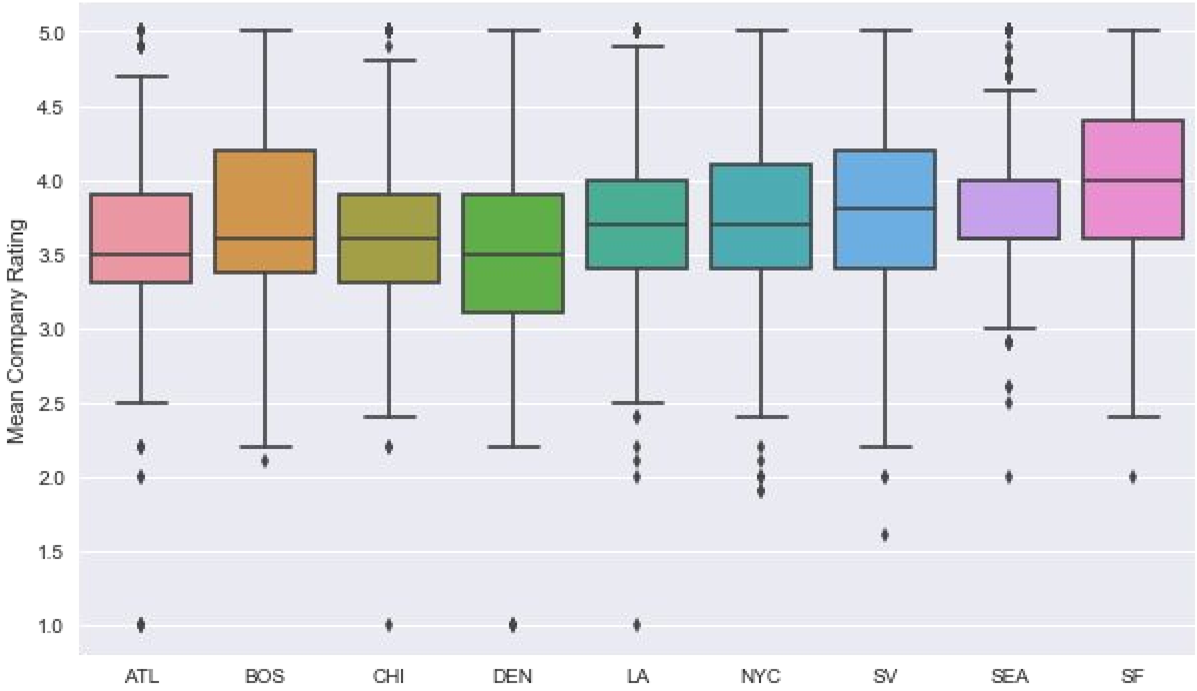
Where to Work: Salary Distribution



Where to Work: Posting Count

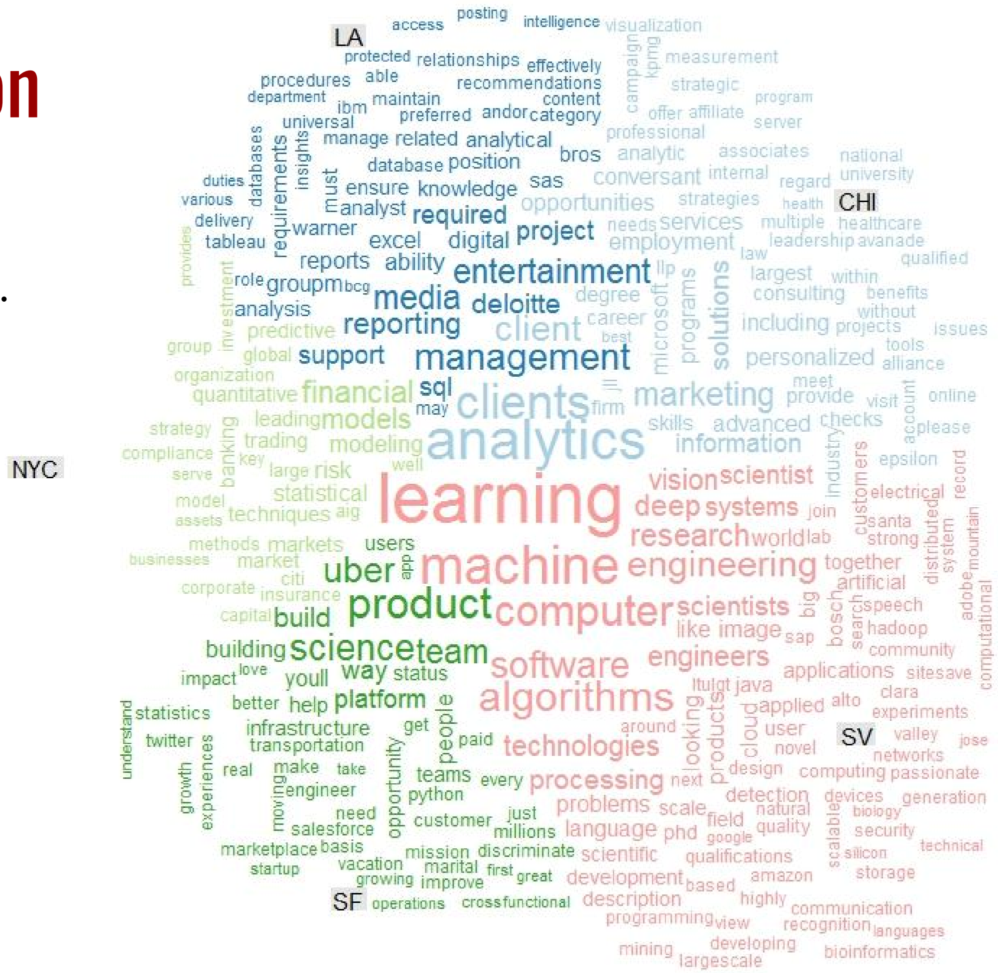


Where to Work: Distribution of Company Ratings



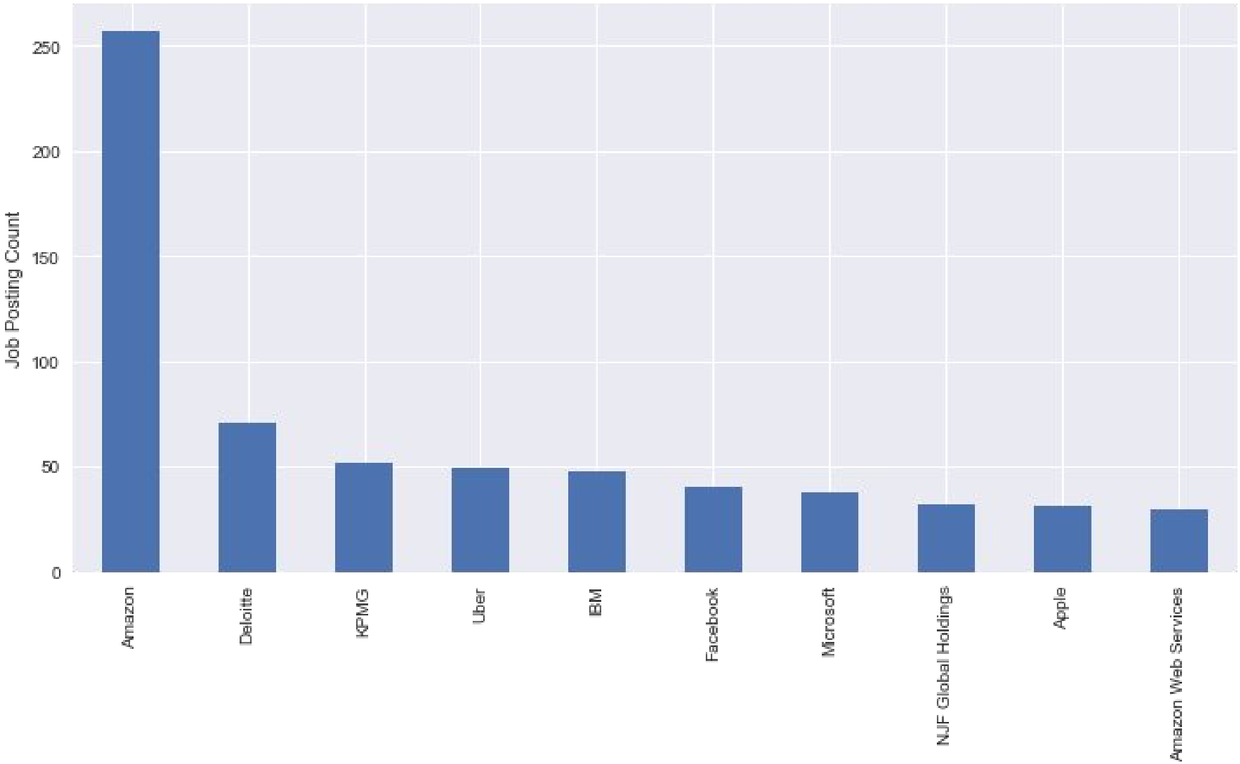
Region: Term Comparison

- LA: entertainment, digital media, consulting, advertising.
- NYC: compliance, banking, trading, risk management.
- SF: Uber, tech industry.
- SV: tech industry (machine learning, AI, NLP)
- CHI: strategy consulting, marketing,



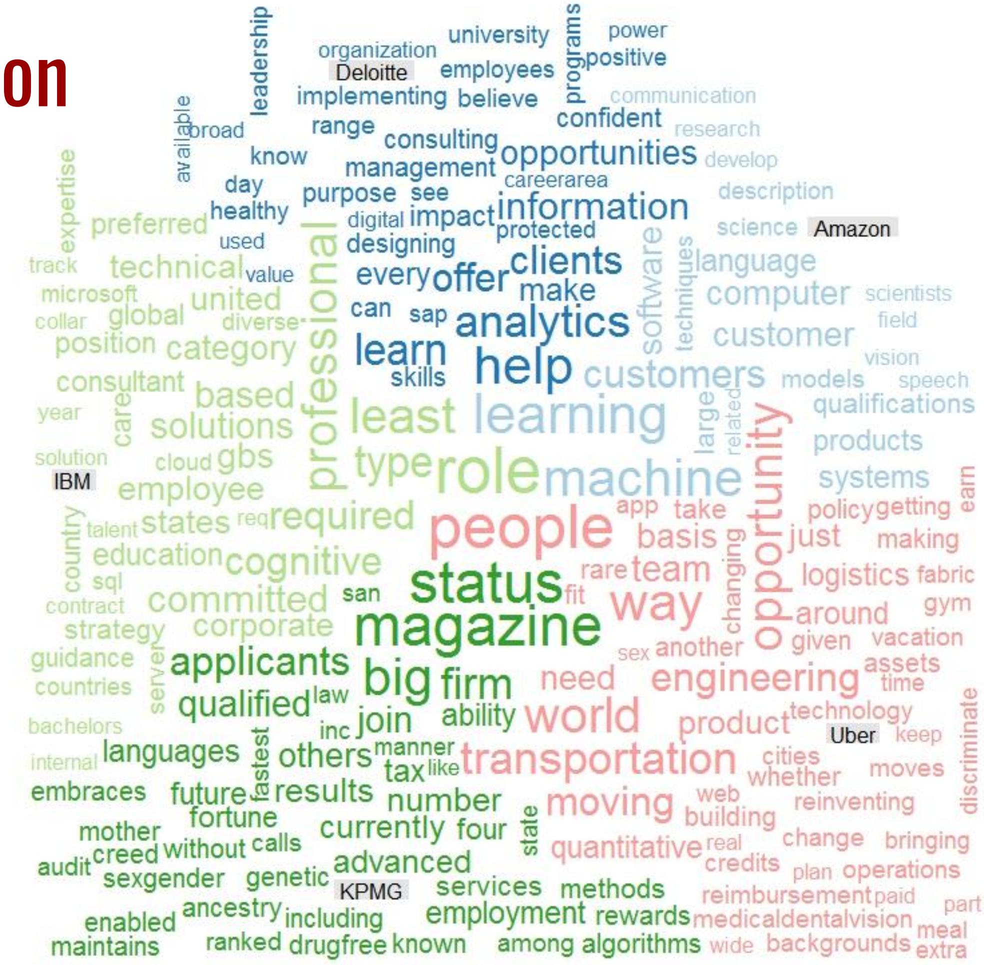
Who is Hiring?

Who is Hiring: Posting Count by Company



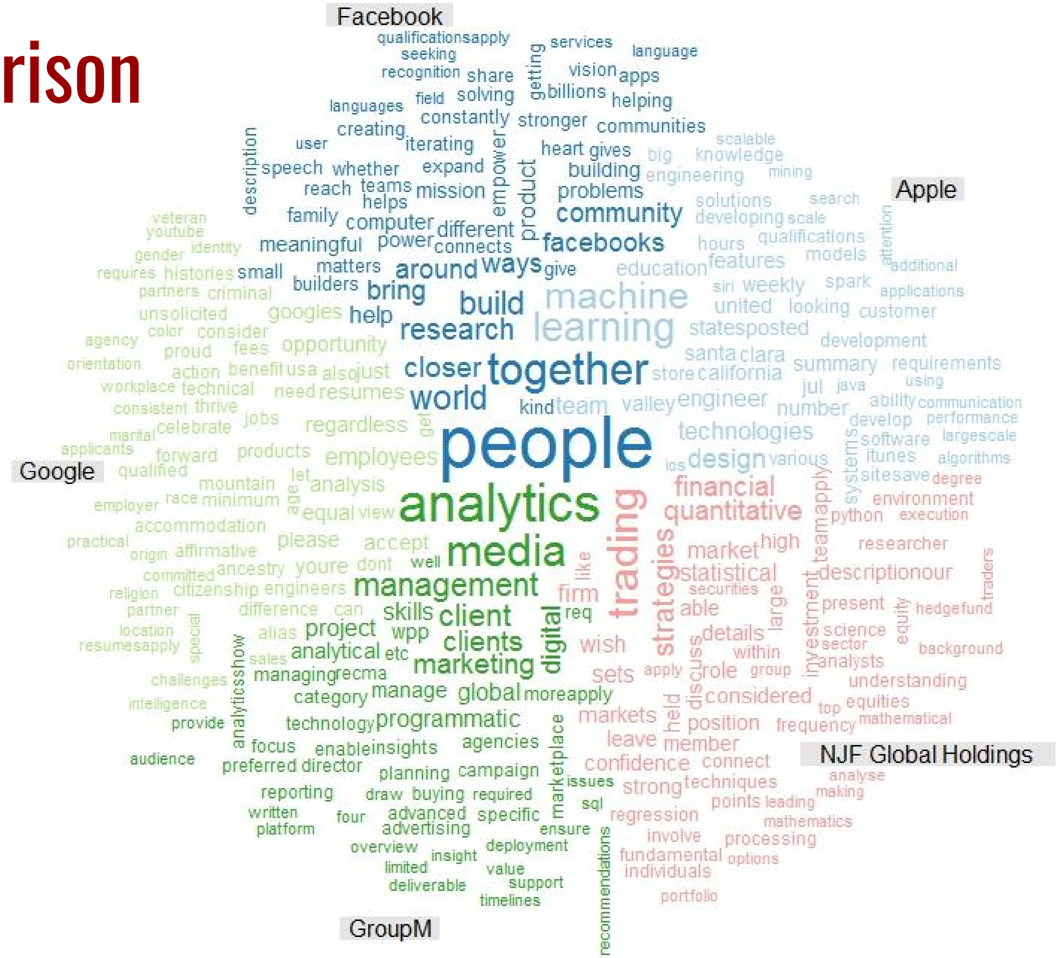
Company: Term Comparison

- Each word is shown under the company that used it the most (relative frequency).



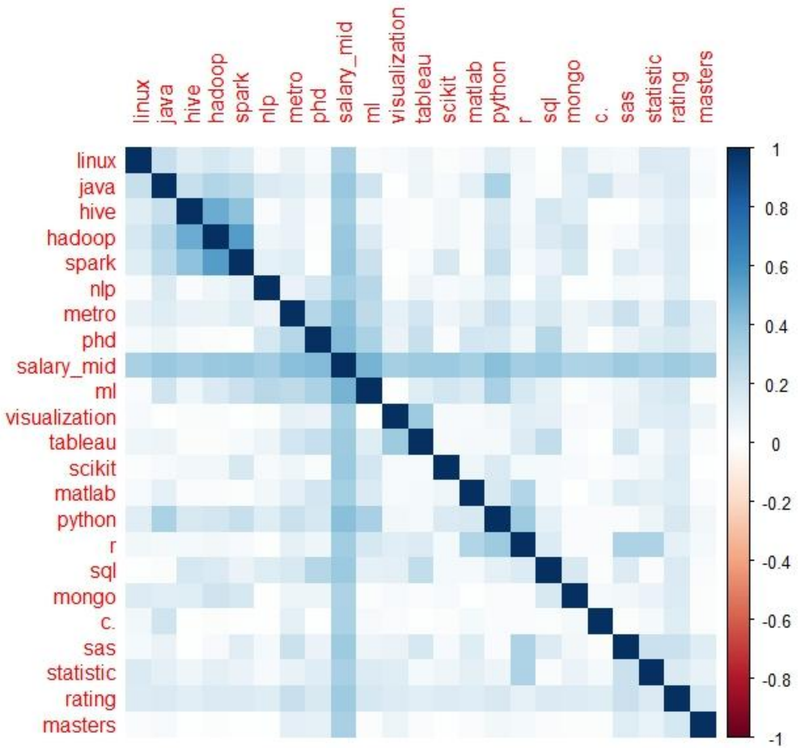
Company: Term Comparison

- Each word is shown under the company that used it the most (relative frequency).



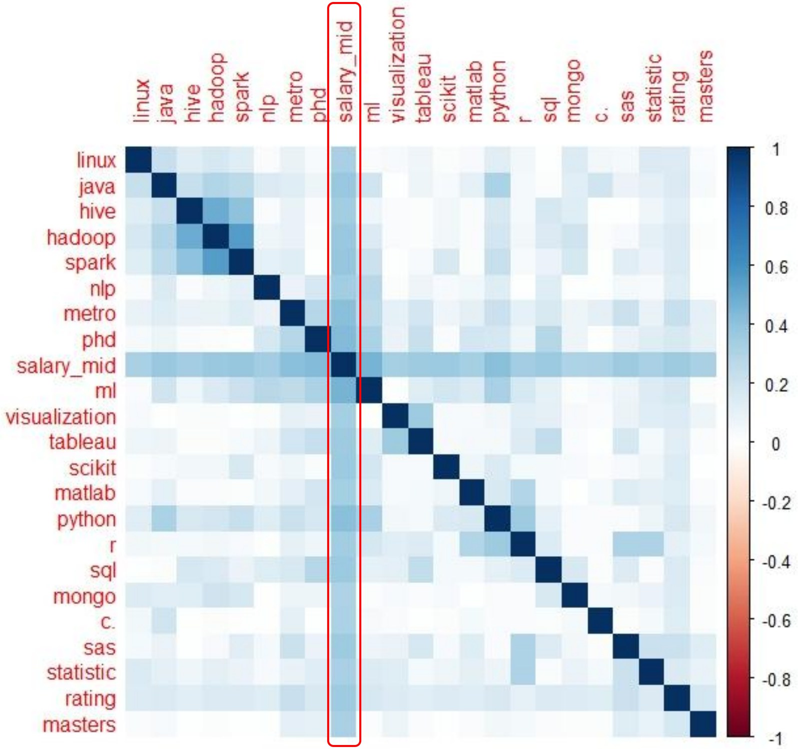
Relationships Between the Variables

- How are the boolean “skills” columns related?
- CorrPlot suggests correlation is near zero or modestly positive for most variables.



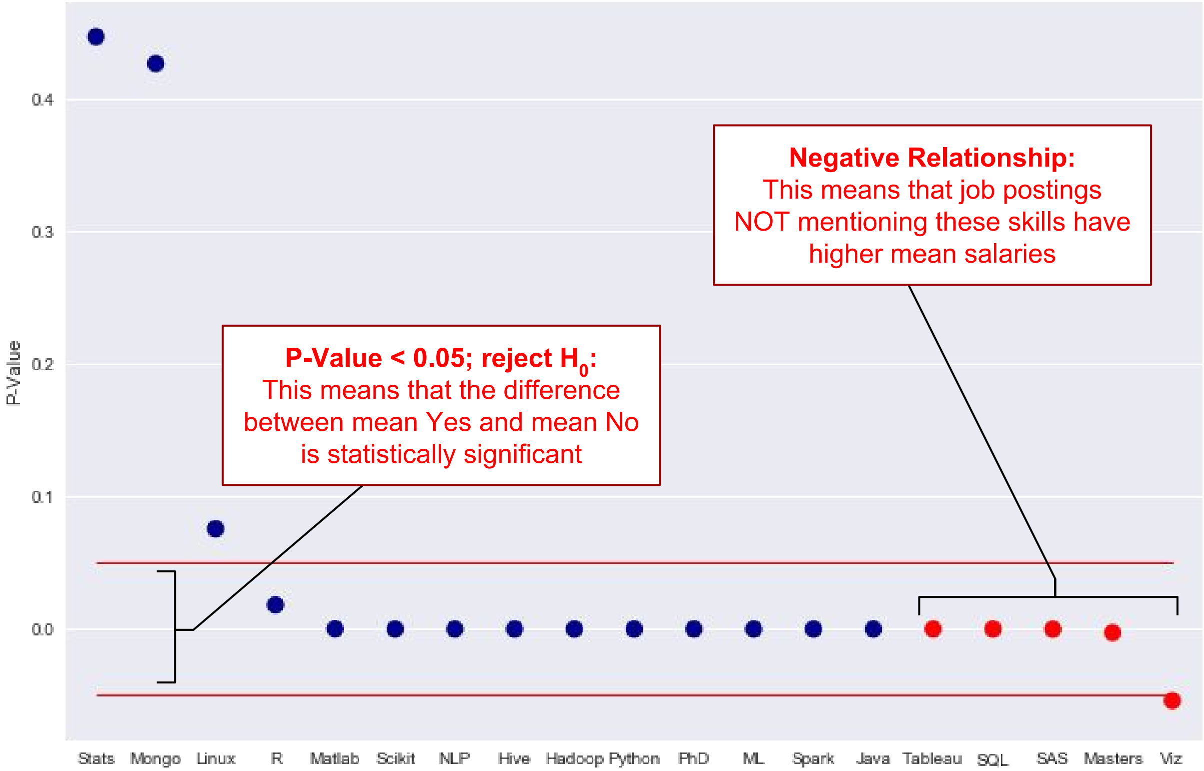
Relationships Between the Variables

- How are the boolean “skills” columns related?
- CorrPlot suggests correlation is near zero or modestly positive for most variables.
- Salary appears positively correlated with all.



Two-Sample T-Tests: Significance of P-Values

- Assumes sample SDs are equal:
  - Largest difference is NLP.
  - 32.9 for Y, 37.0 for Y.
- Assumes population of is normally distributed.
  - Uncertain without delving deeper into the overall job market
- Assumes samples are randomly drawn and independent.
  - Also uncertain, without knowing more about how Glassdoor assigns estimated salaries to job postings.



Negative Relationship:
This means that job postings NOT mentioning these skills have higher mean salaries

P-Value < 0.05; reject H_0:
This means that the difference between mean Yes and mean No is statistically significant

Conclusions

Conclusions

- Learn Python, R, Scikit, Machine Learning, NLP, Java, Spark....
    - (and maybe get a PhD)
- SAS, Tableau, SQL, and Master's degrees tend to be associated with lower salaries.
- Salaries tend to be higher in Silicon Valley, San Francisco, and Seattle.

Conclusions

- Learn Python, R, Scikit, Machine Learning, NLP, Java, Spark....
  - (and maybe get a PhD)
- SAS, Tableau, SQL, and Master's degrees tend to be associated with lower salaries.
- Salaries tend to be higher in Silicon Valley, San Francisco, and Seattle.
- Opportunities for future work:
  - Multiple linear regression: test predictive power of variables on salary.
  - Sentiment analysis: search for differences in tone of job description text by location, title, or company.
  - Topic modeling: use TD/IDF and cosine similarity to do a more detailed analysis on the similarities and differences between postings.
  - Recommendation engine: allow users to select skills, and return suitable job postings.