Scraping Glassdoor:
automating your job search

Aungshuman Zaman
NYC Data Science Academy
August 8, 2018

Motivation of analysis

- Instead of relying on hearsay wanted to explore job market and gather information.
- But I am lazy and want to automate.
- Wanted to answer the following questions:
  - How many data science/ data engineer or data analysis jobs are out there?
  - What is the salary?
  - Does location matter?
  - What qualifications do you need?
  - What skills are most valuable?

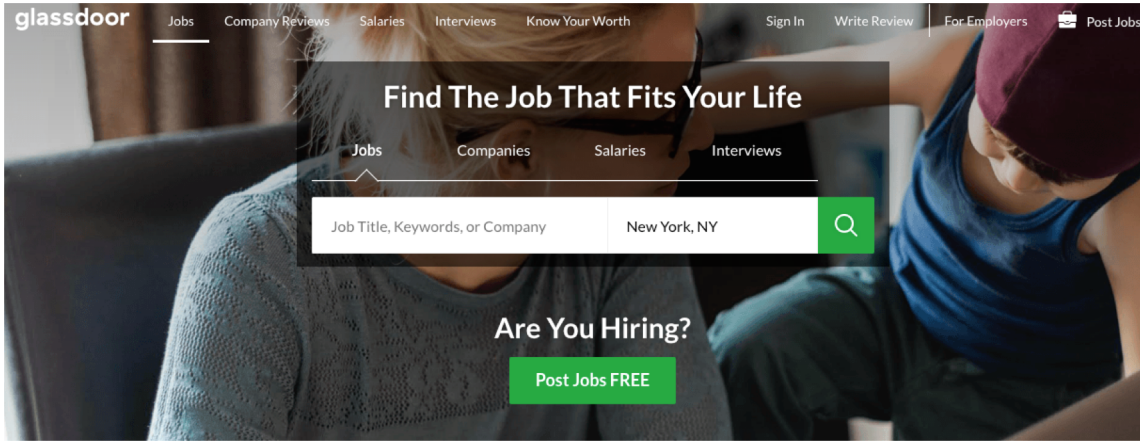
glassdoor®

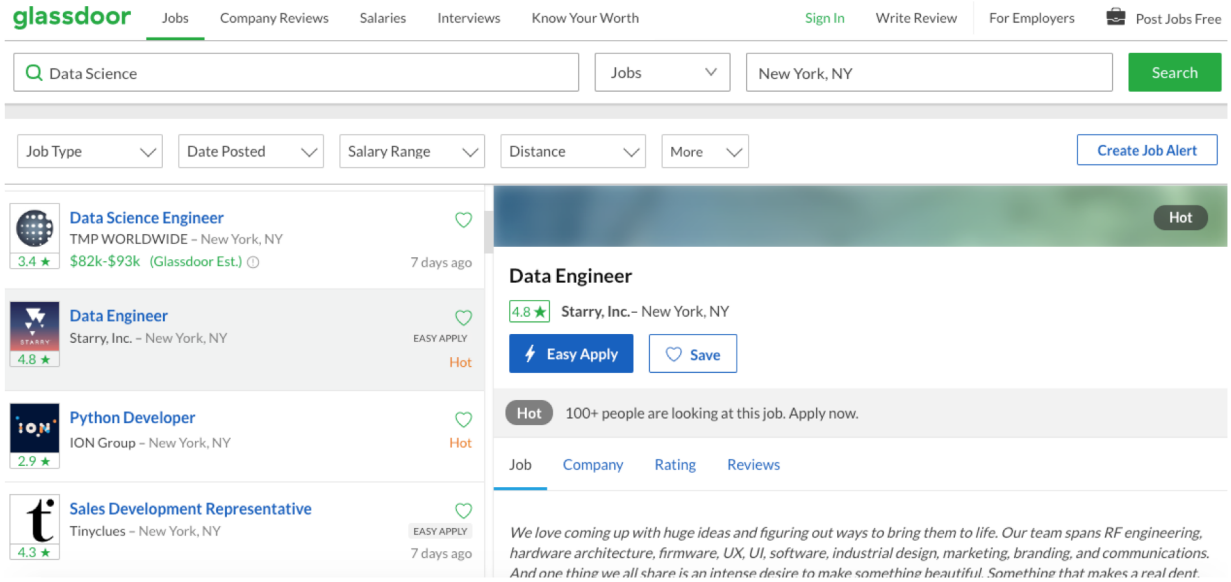
- Decided to scrape the glassdoor website to get my data. Other options were indeed or monster.
- An American company established in 2007
- Originally the idea was to collect company reviews and real salaries from employees of large companies and display them anonymously.
- Now contains millions of job posts, salary estimates and reviews.

Scraping the website

Used Selenium

Glassdoor only allows 30 or so pages for each query





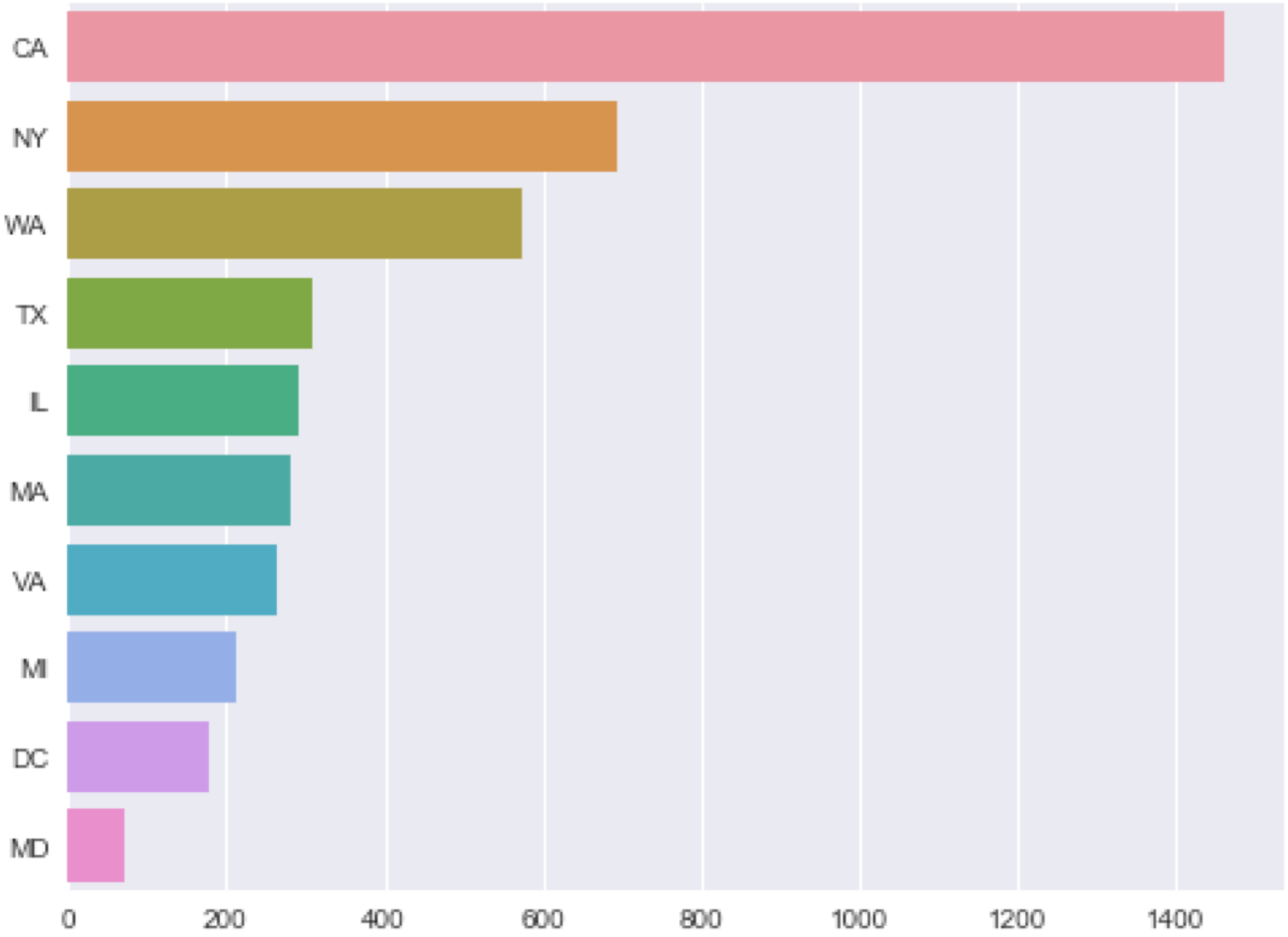
Decided to scrape for data science/ engineer/ analyst jobs in different US cities.

The scraped dataset

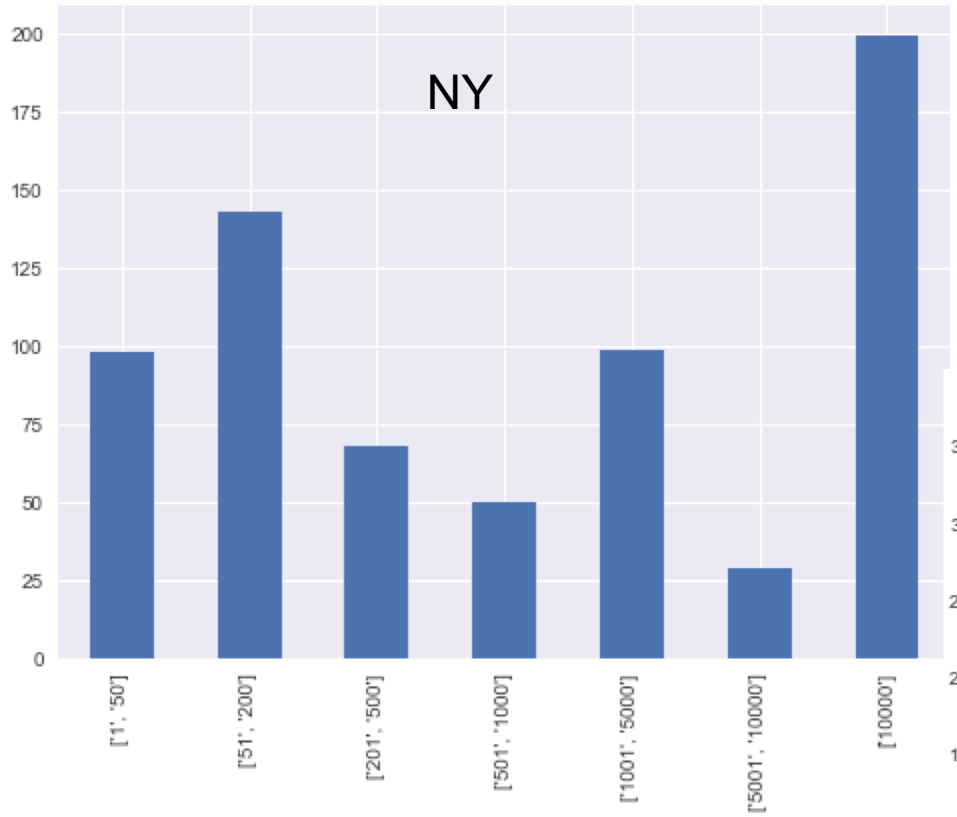
- I scraped in total 4500 entries across US.
- Information about
  - Job position
  - Company information
    - Where it is based
    - Size
    - Industry
    - Rating
  - City and state where job is
  - Salary range (glassdoor estimates this from reports from employees)
  - Description of job posting (I create a bag of word from each entry)

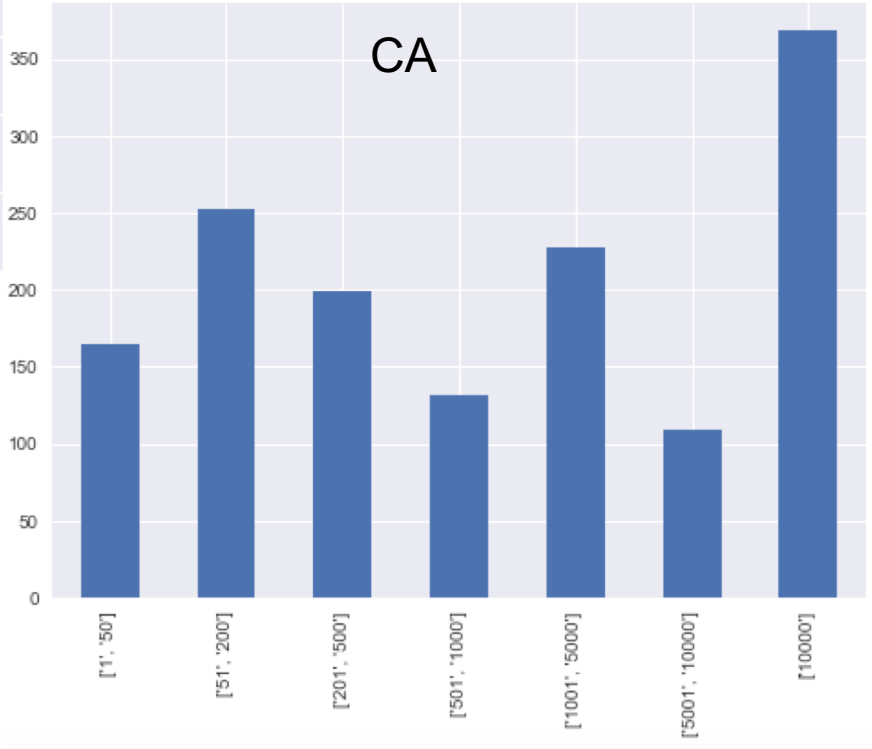
State-wise job posting*

* I looked into the largest cities. Sample is not unbiased, so difficult to say much about population.



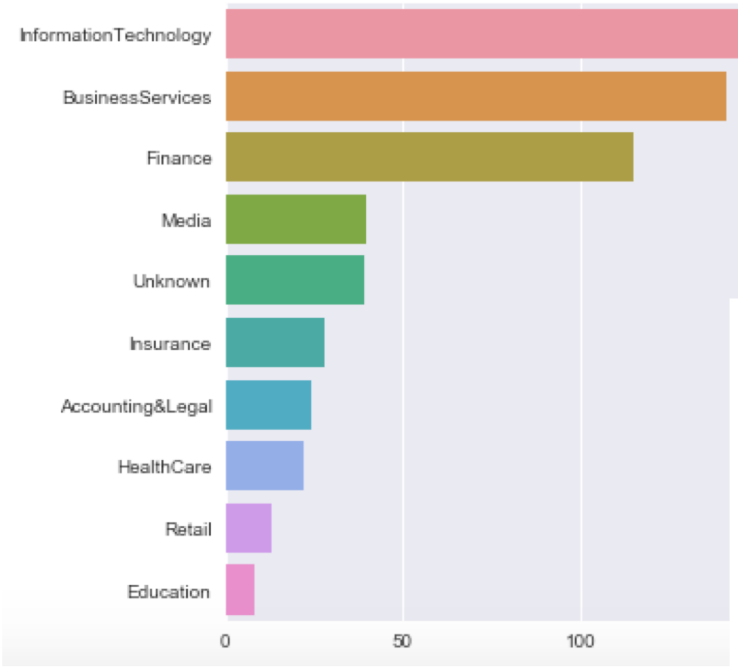
Jobs vs Company size in NY and CA



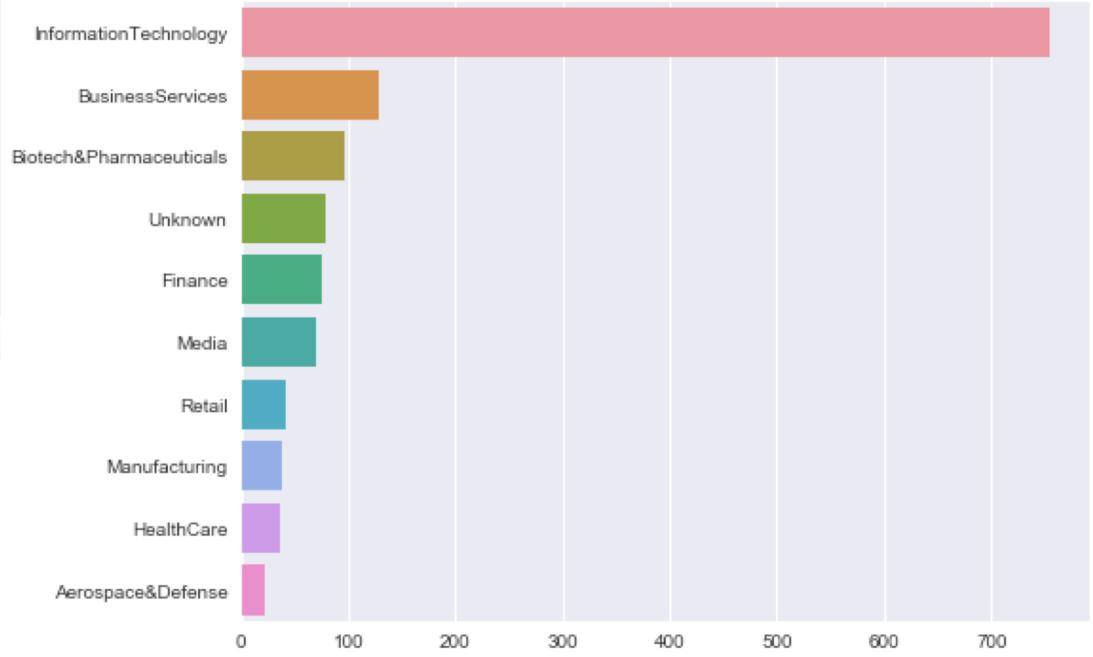


NY: which industry offers most jobs?

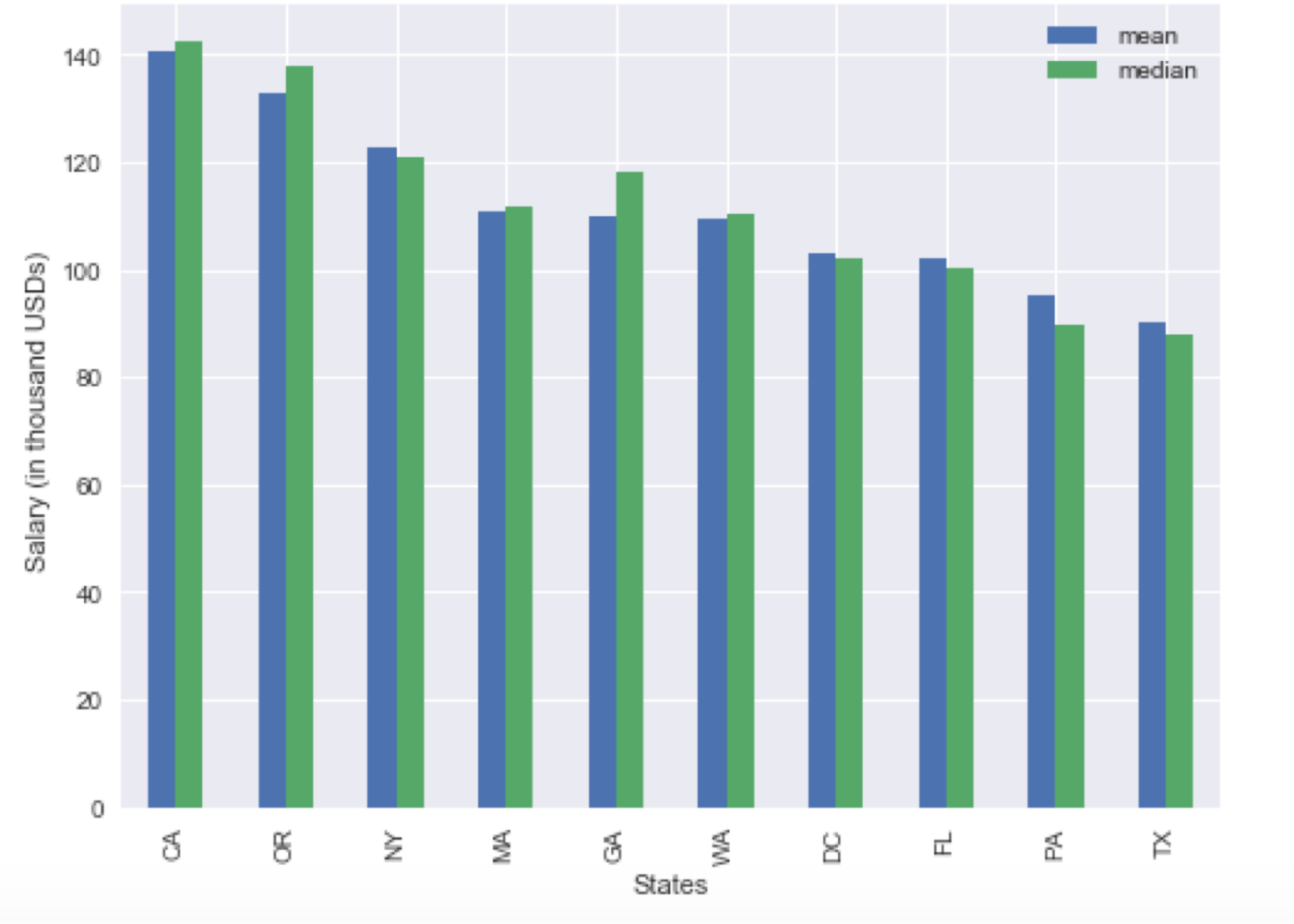
NY



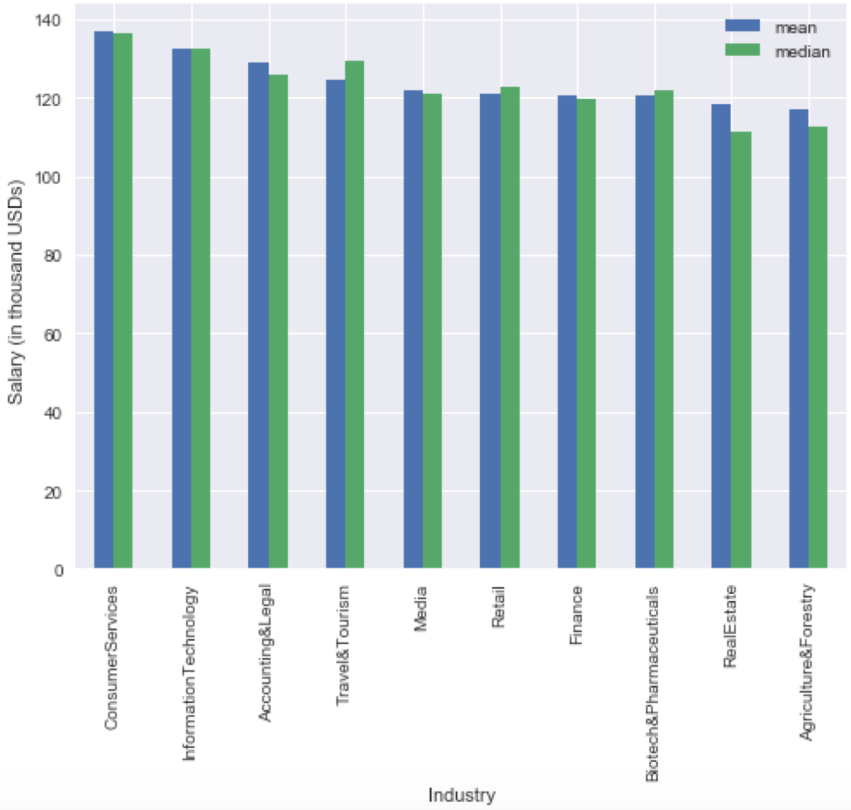
CA

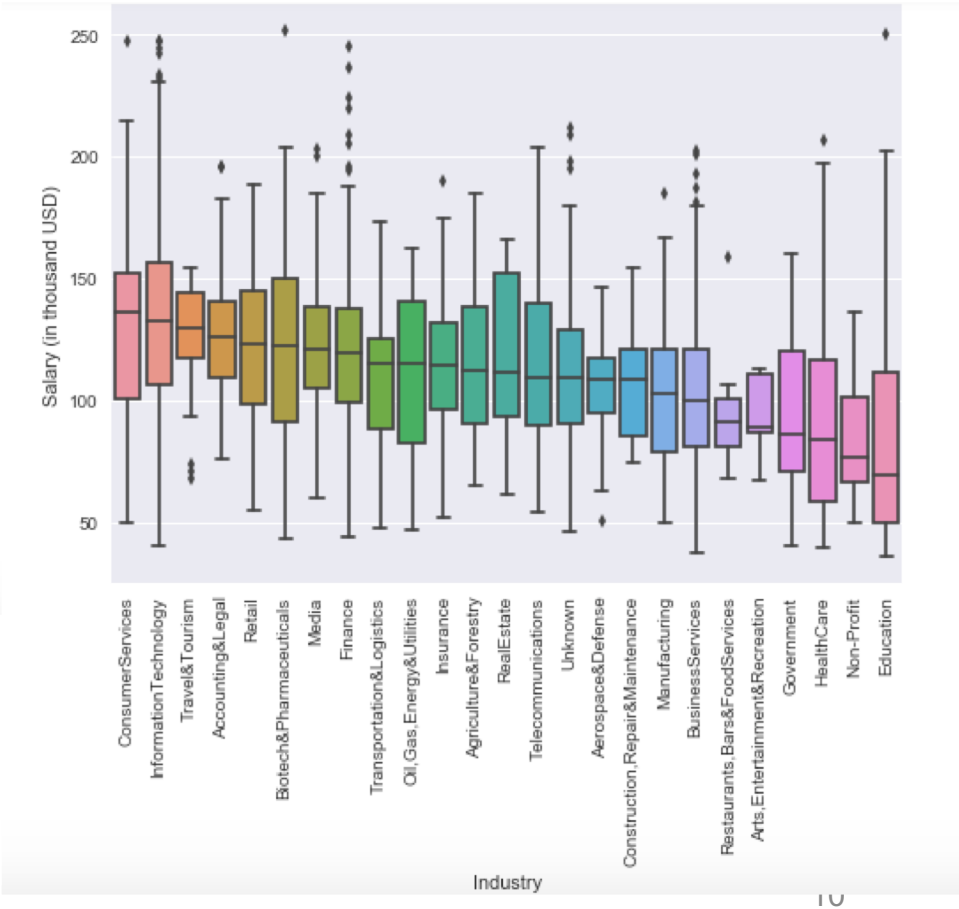


Mean and median salary by states

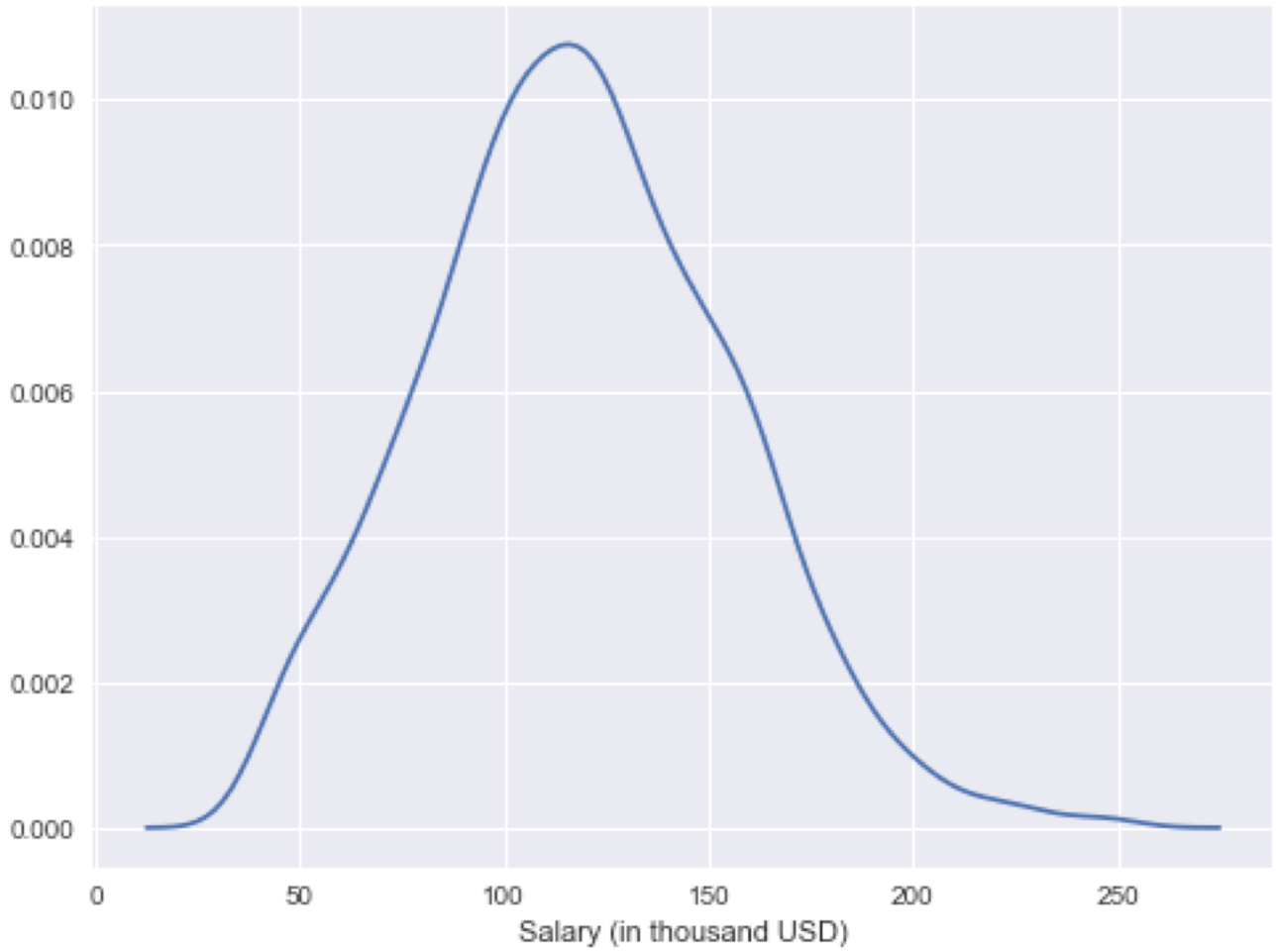


Salary by industry



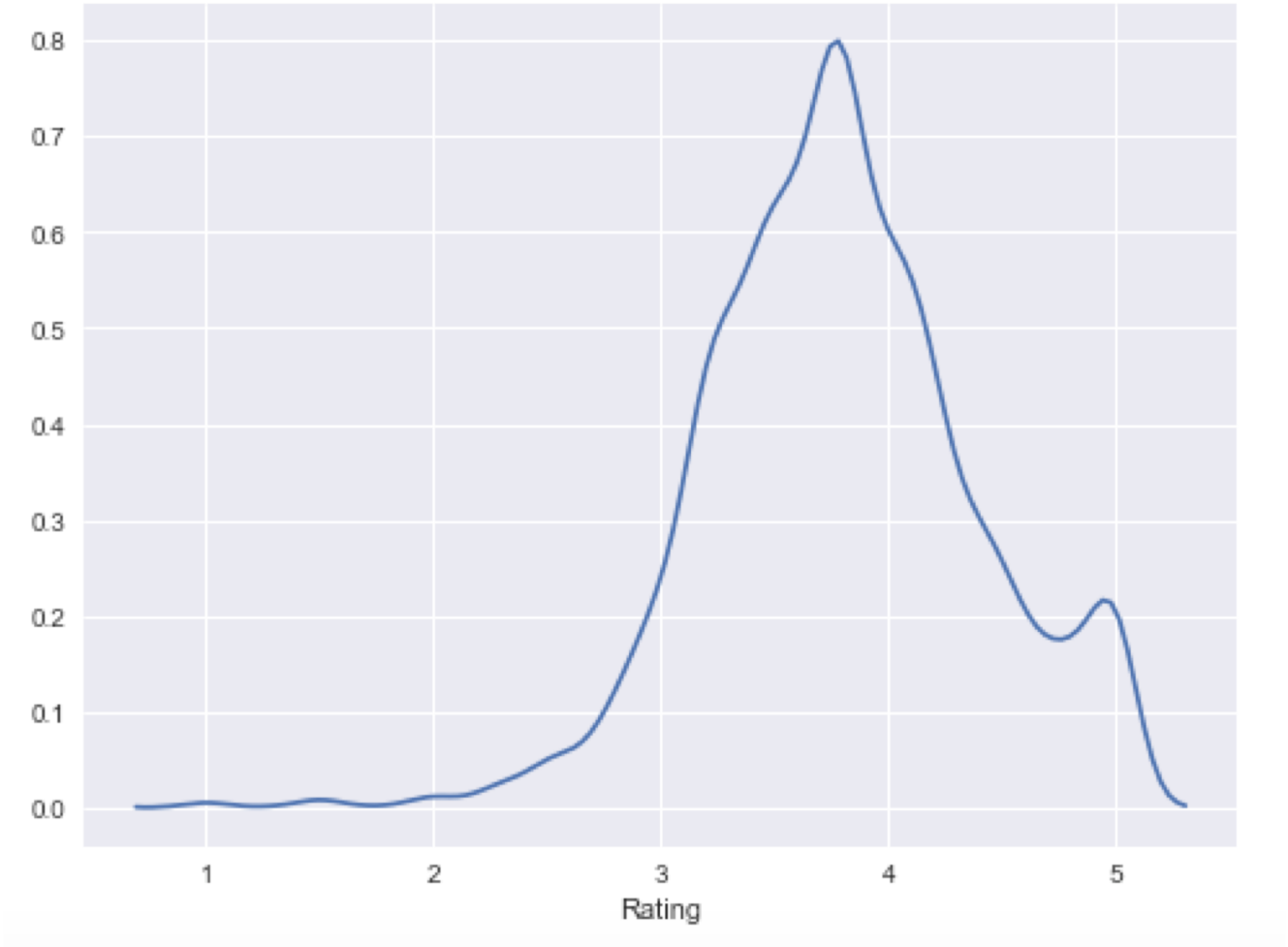


Overall salary distribution*

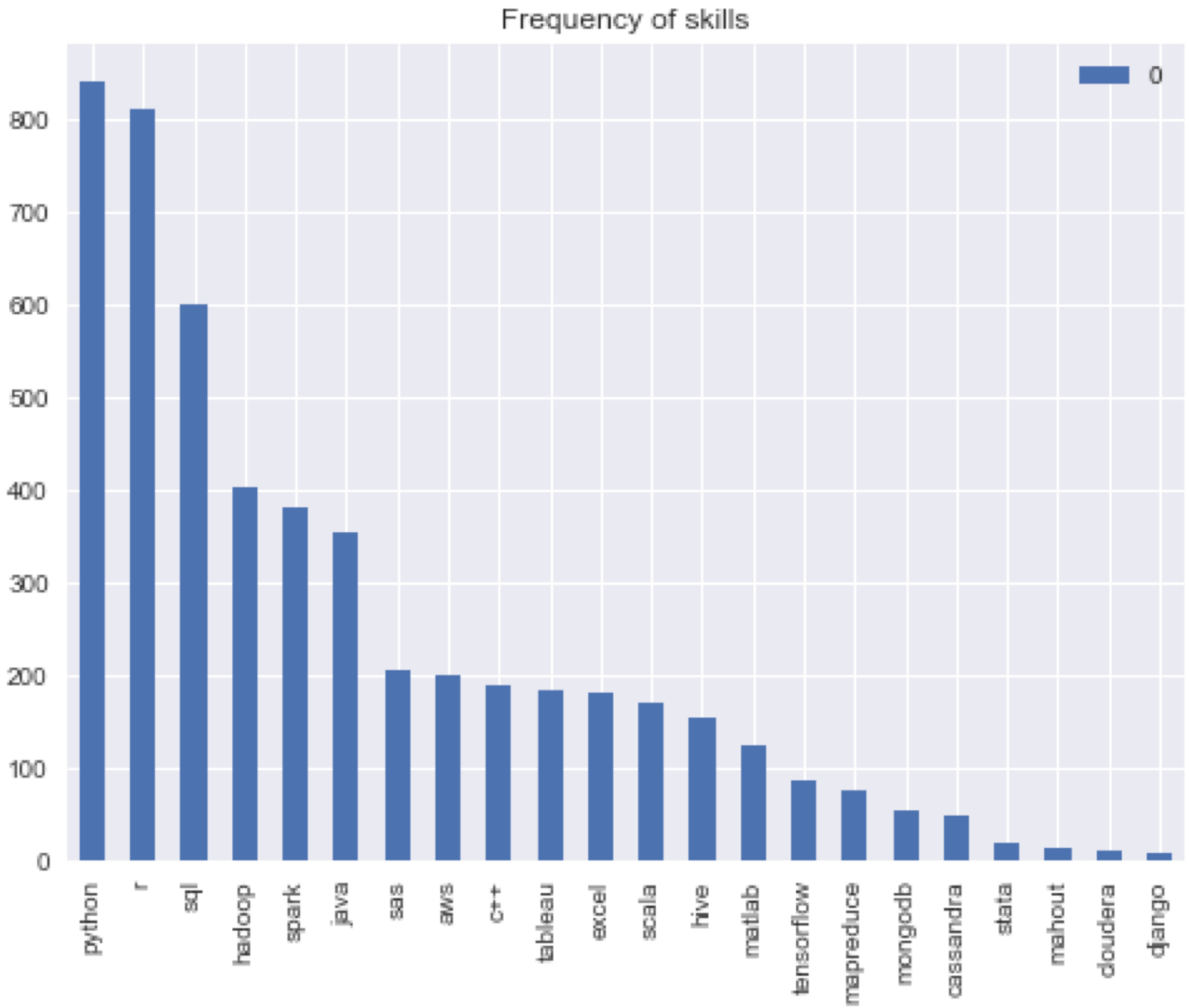


*No distinction between job types, or location.

Company rating distribution



Important skill-sets for data related jobs



Education level requirements

Word	Number of posts it appears on
phd, doctorate, doctor	374
masters, ms:	590
bachelor, bs, undergraduate, associate	528

Future analysis?

- More granular look at job posting--> differentiate between data science/ engineer/ analyst
- Compare with work on other job sites.
- Explore job qualification, factors like experience.
- Given your skill sets, can we recommend which jobs you should apply to?
- Given location, salary, company size, industry etc. can we predict job satisfaction (rating)?

Thank You!