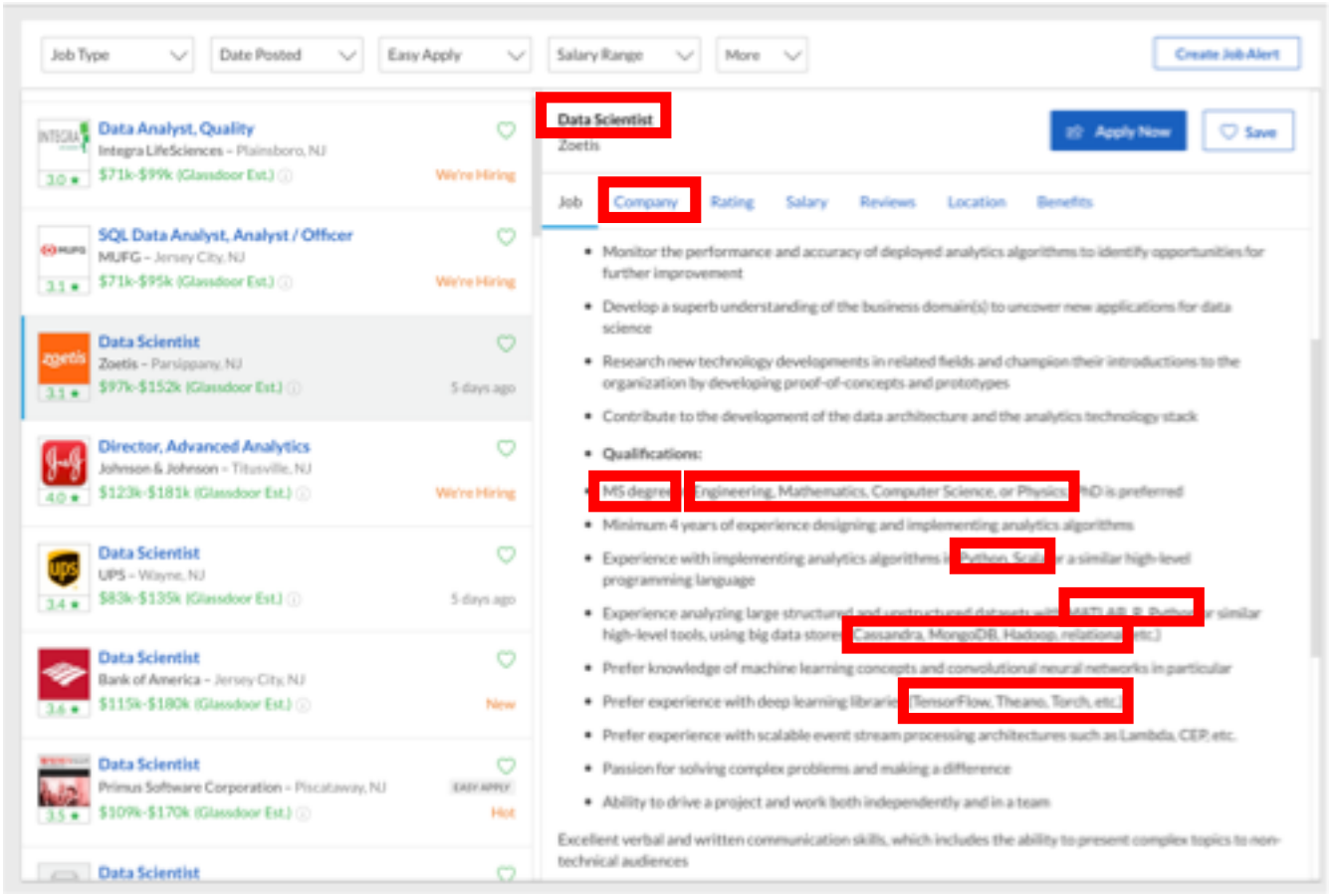
Scrapping Glassdoor

Rajesh Arasada
NYC Data Science Academy 2018

The Questions

- Skills companies were looking for when hiring
    - Data Analyst
    - Data Scientist
    - Data Engineer
- What companies are actively hiring?
- Does the skill requirement vary with the type of industry?

The Scrapping Process



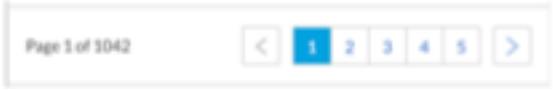
The Scrapping Process

Specific URL for every State

'https://www.glassdoor.com/Job/maine-data-science-jobs-SRCH_IL.0,5_IS758_KO6,18_IP1.htm'

'https://www.glassdoor.com/Job/ohio-data-science-jobs-SRCH_IL.0,4_IS2235_KO5,17_IP1.htm'

URL changes between pages



URLs do not change between job listings

Obtained job IDs with Scrapy

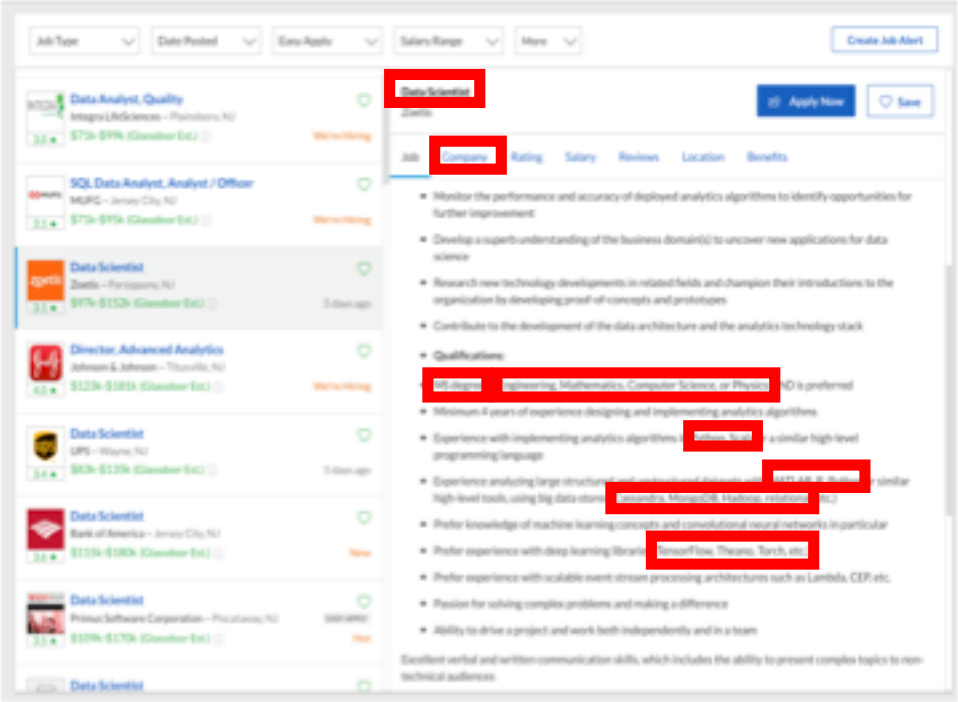
Used Selenium to extract the data

Specific job Listing IDs

jobListingId=2822327760',
jobListingId=2795371564',

Parsing Data

- Job Description
- Job Title
- Address
- Industry
- Sector



Job Description

- Regular Expression
- Key words

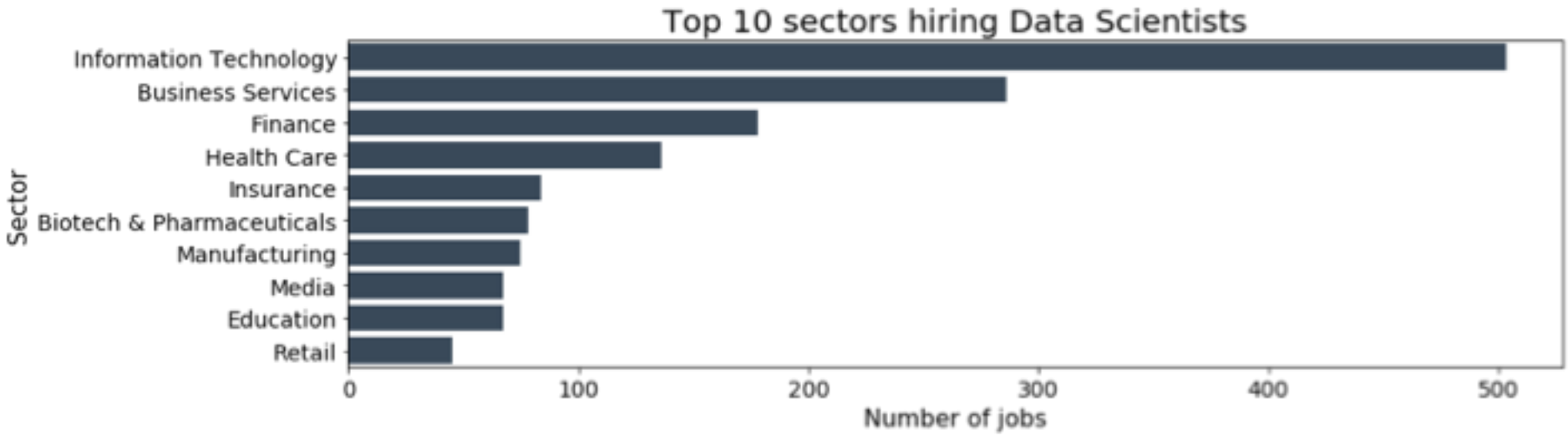
Job Titles

- 250 unique job titles in NY alone
- Binned them into:
  - Analyst (Lead, Intern, Junior)
  - Scientist
  - Engineer
  - Others (Managers, Director)

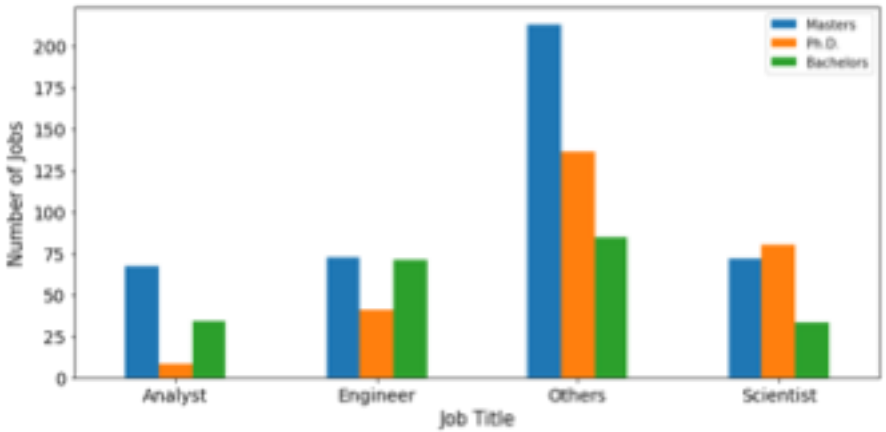
Address

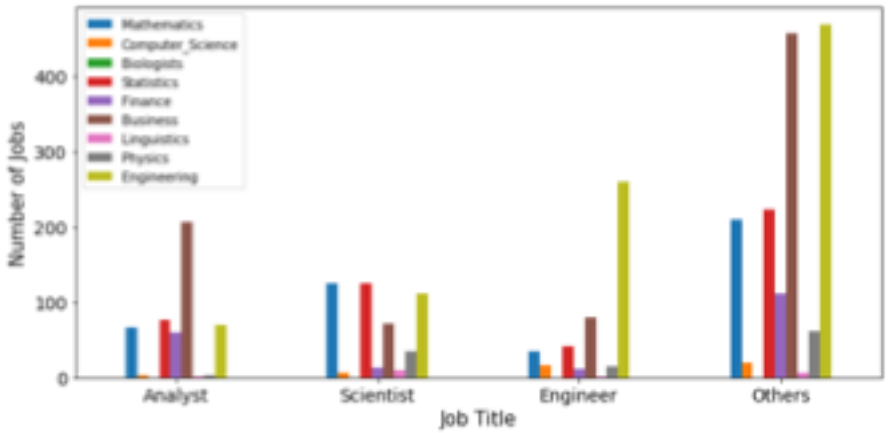
- Managed to scrape data from 6 states: MA, NY, NJ, PA, TX, CA

Which industry sectors are hiring data scientists?

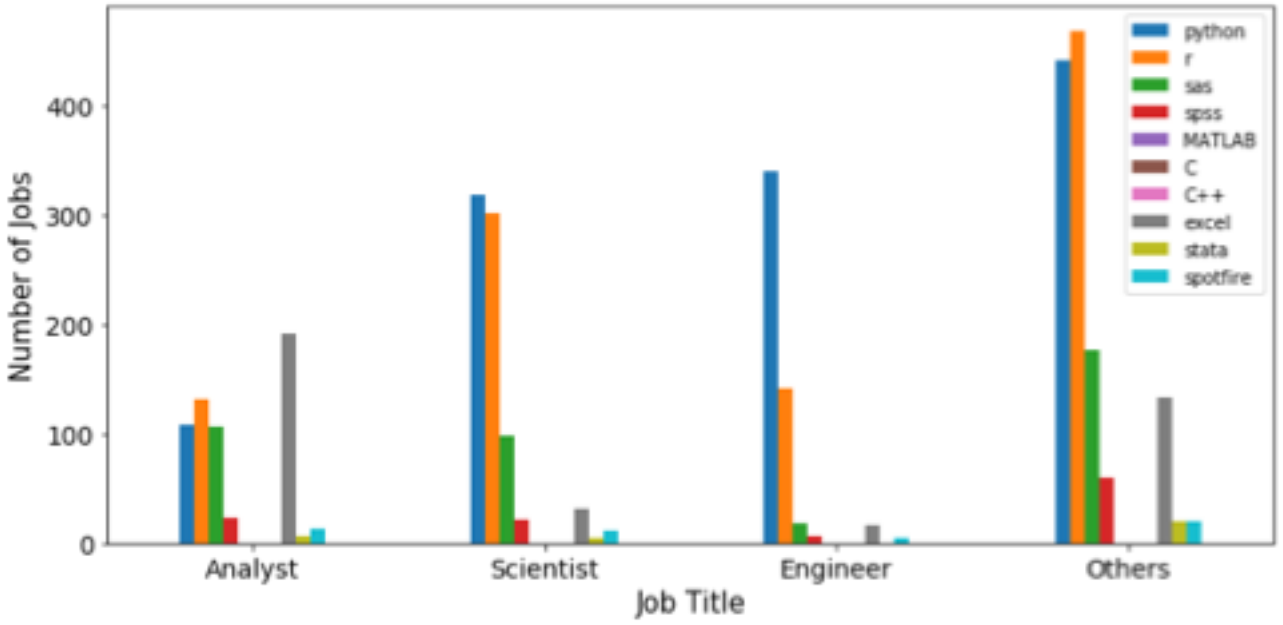


What are the educational requirements for data science jobs?

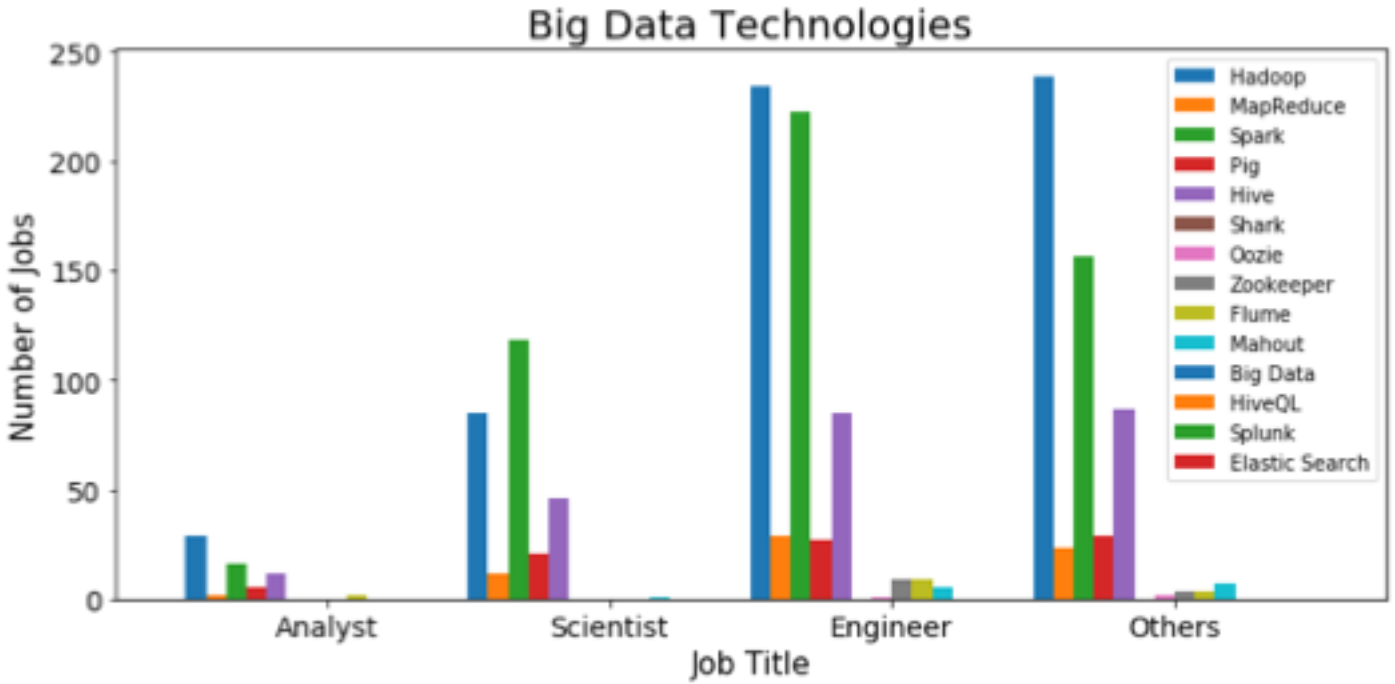




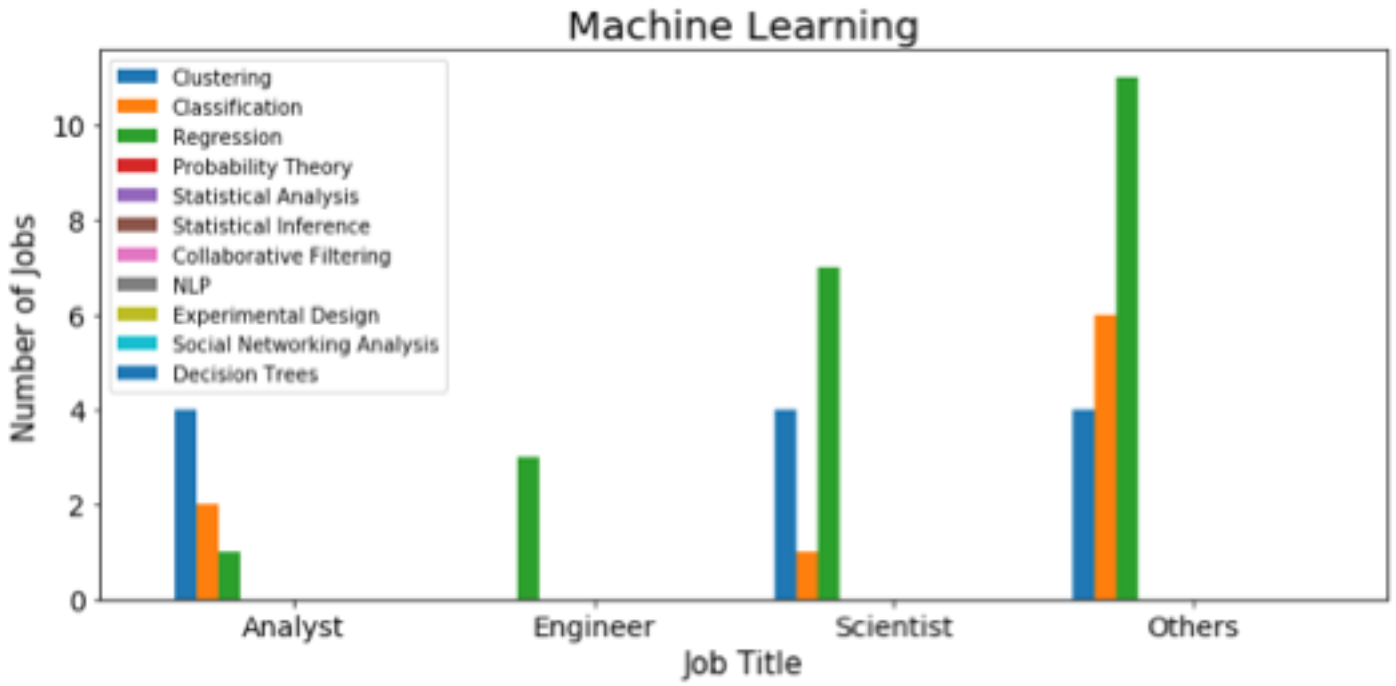
What are the important language skills for the data science jobs?



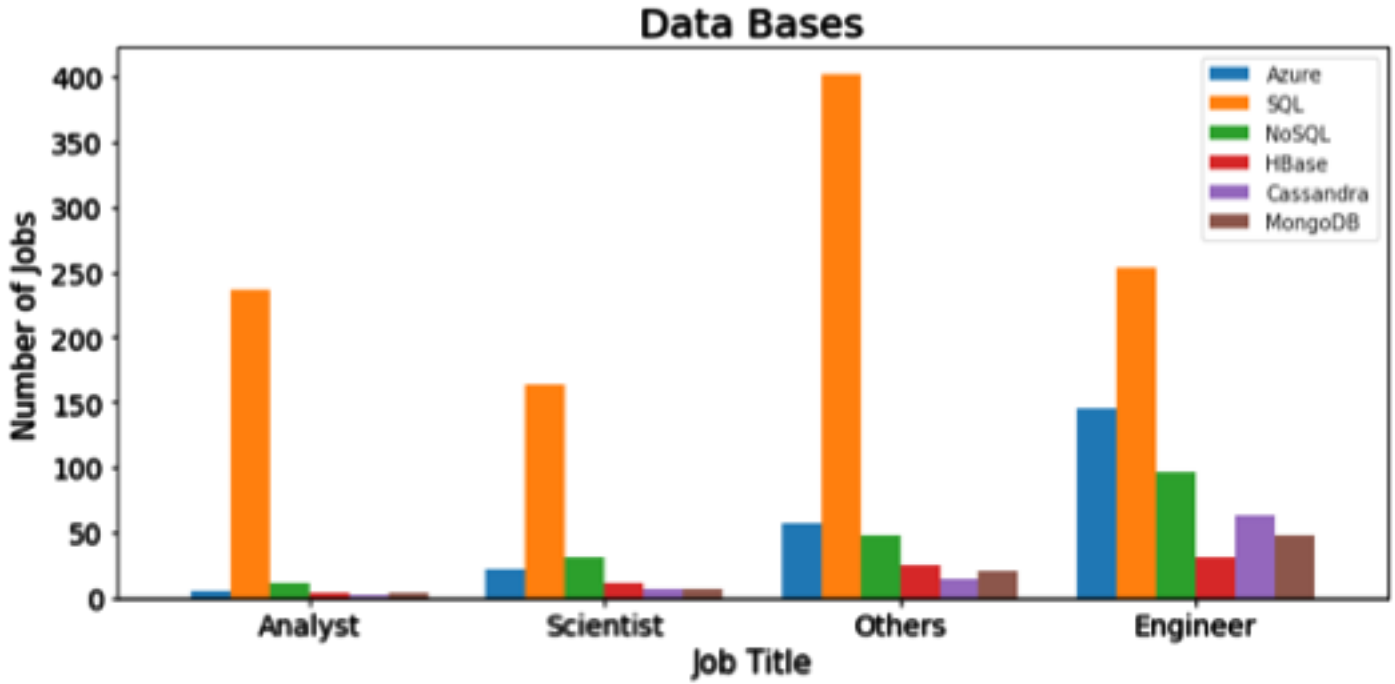
What are the important skills for the data science jobs?



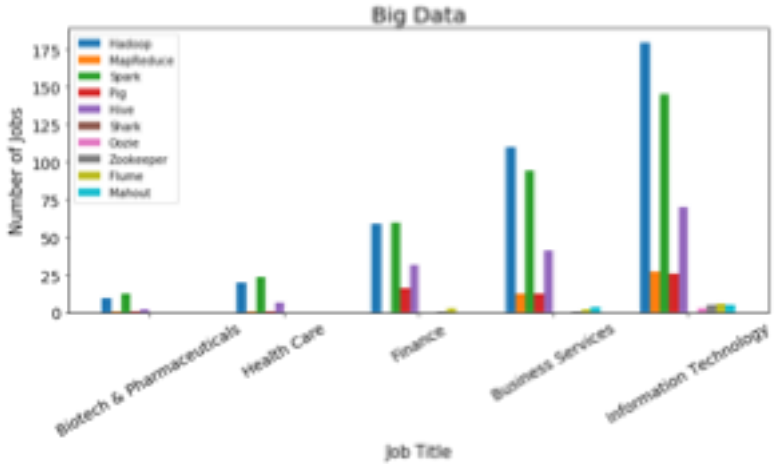
What are the important skills for the data science jobs?

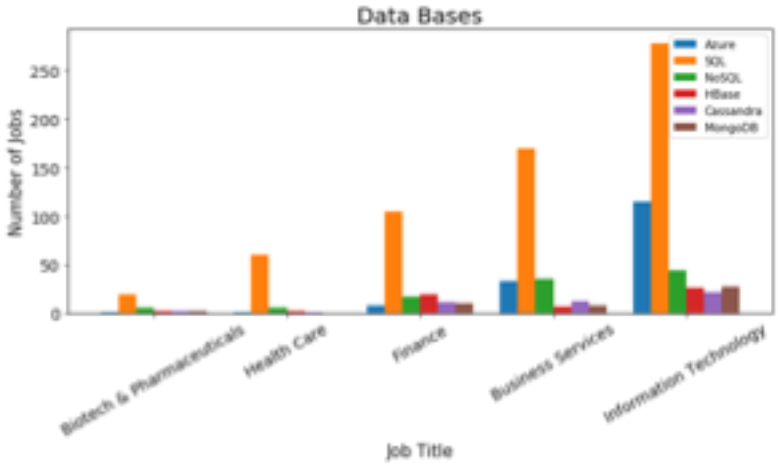


What are the important skills for the data science jobs?

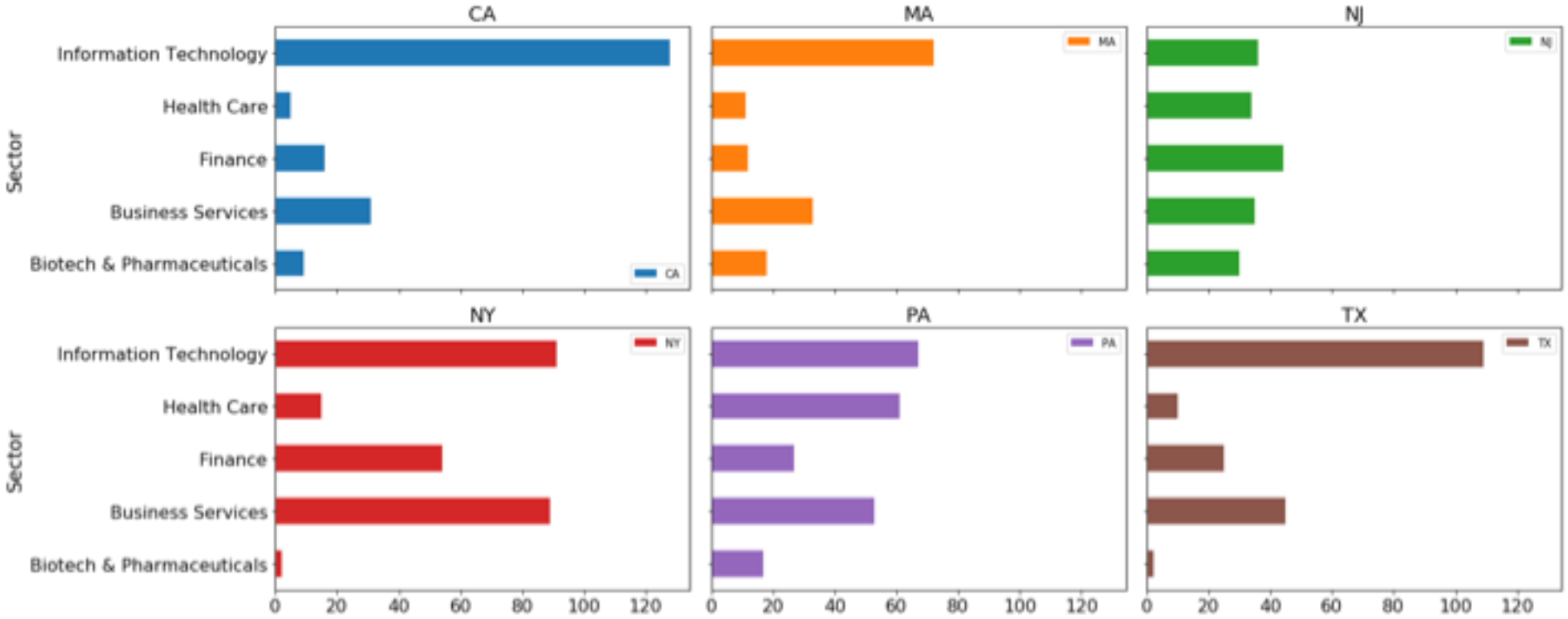


Does the skill requirements vary with the industry sector?





Distribution of data science jobs across states and sectors



Further Analysis

- Collect more data across different states and cities
- Improve the code to identify the important words efficiently
- Improve on the classification of job titles

Thank you