

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 screenshot displays a job search interface. At the top, there are filters for Job Type, Date Posted, Easy Apply, Salary Range, and More. A 'Create Job Alert' button is also present. The main list of jobs includes:

- Data Analyst, Quality** at Integra LifeSciences - Plainsboro, NJ. Rating: 3.0. Salary: \$71k-\$99k (Glassdoor Est.). Status: We're Hiring.
- SQL Data Analyst, Analyst / Officer** at MUFG - Jersey City, NJ. Rating: 3.1. Salary: \$71k-\$95k (Glassdoor Est.). Status: We're Hiring.
- Data Scientist** at Zoetis - Parsippany, NJ. Rating: 3.1. Salary: \$97k-\$152k (Glassdoor Est.). Status: 5 days ago.
- Director, Advanced Analytics** at Johnson & Johnson - Titusville, NJ. Rating: 4.0. Salary: \$123k-\$181k (Glassdoor Est.). Status: We're Hiring.
- Data Scientist** at UPS - Wayne, NJ. Rating: 3.4. Salary: \$83k-\$135k (Glassdoor Est.). Status: 5 days ago.
- Data Scientist** at Bank of America - Jersey City, NJ. Rating: 3.6. Salary: \$115k-\$180k (Glassdoor Est.). Status: New.
- Data Scientist** at Primus Software Corporation - Piscataway, NJ. Rating: 3.5. Salary: \$109k-\$170k (Glassdoor Est.). Status: Hot.

The detailed view of the **Data Scientist** role at Zoetis is shown on the right. The job title is highlighted in a red box. The job description includes:

- Qualifications:**
 - MS degree in Engineering, Mathematics, Computer Science, or Physics. PhD is preferred.
 - Minimum 4 years of experience designing and implementing analytics algorithms.
 - Experience with implementing analytics algorithms using Python, Scala, or a similar high-level programming language.
 - Experience analyzing large structured and unstructured datasets with MATLAB, R, Python, or similar high-level tools, using big data stores (Cassandra, MongoDB, Hadoop, relational, etc.).
 - Prefer knowledge of machine learning concepts and convolutional neural networks in particular.
 - Prefer experience with deep learning libraries (TensorFlow, Theano, Torch, etc.).
 - Prefer experience with scalable event stream processing architectures such as Lambda, CEP, etc.
 - Passion for solving complex problems and making a difference.
 - Ability to drive a project and work both independently and in a team.
- Excellent verbal and written communication skills, which includes the ability to present complex topics to non-technical audiences.

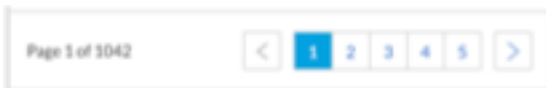
The Scrapping Process

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

Specific job Listing IDs

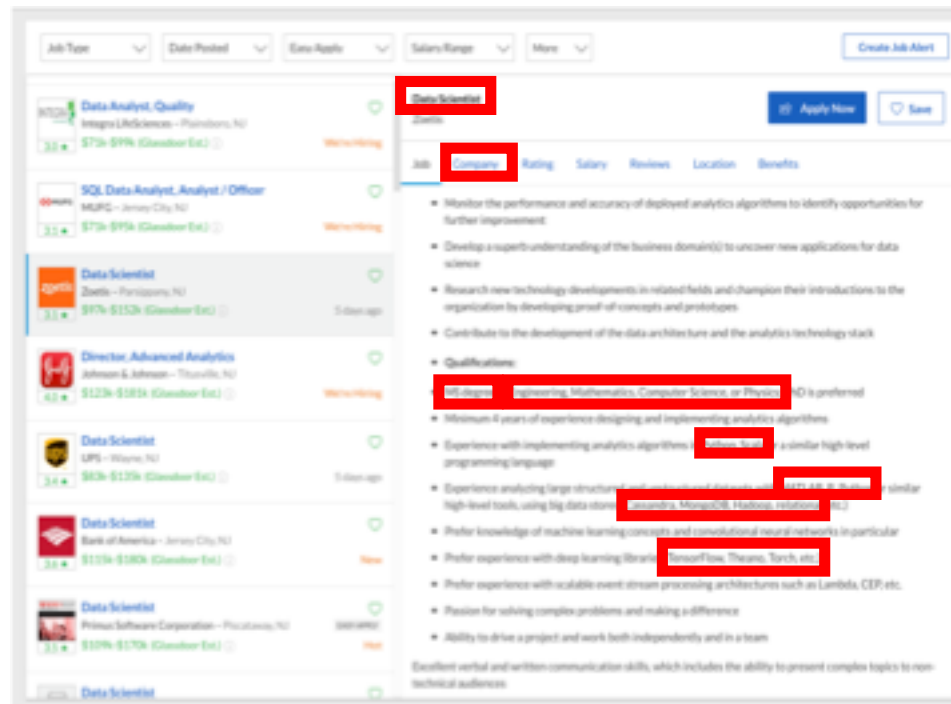
jobListingId=2822327760',

jobListingId=2795371564',

Obtained job IDs with Scrapy
Used Selenium to extract the data

Parsing Data

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



Job Description

- Regular Expression
- Key words

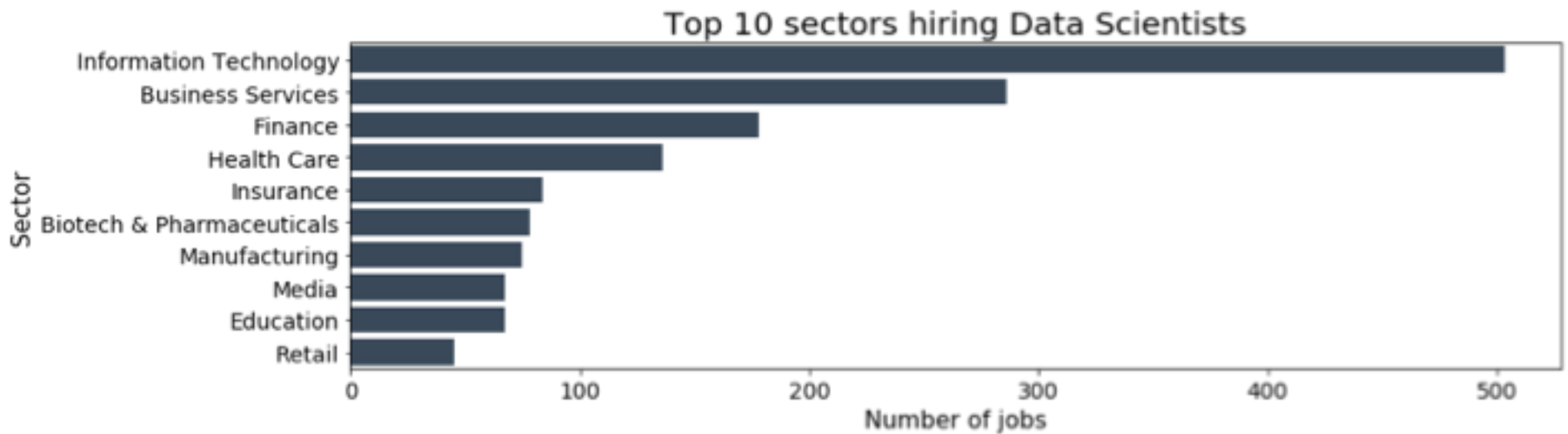
Job Titles

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

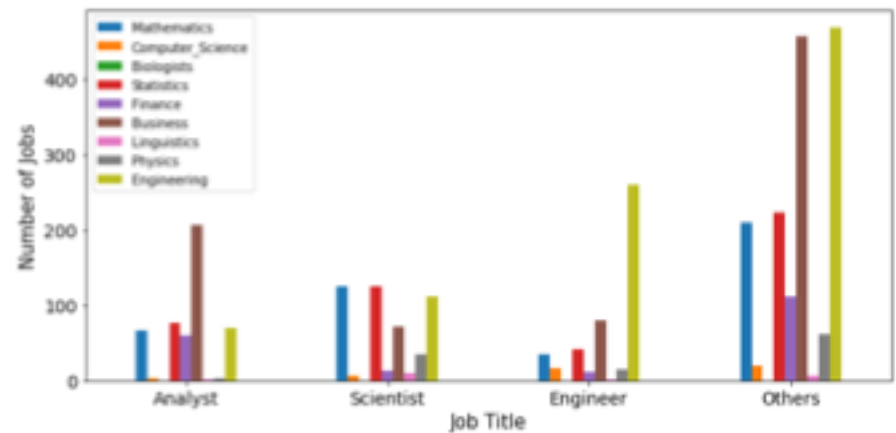
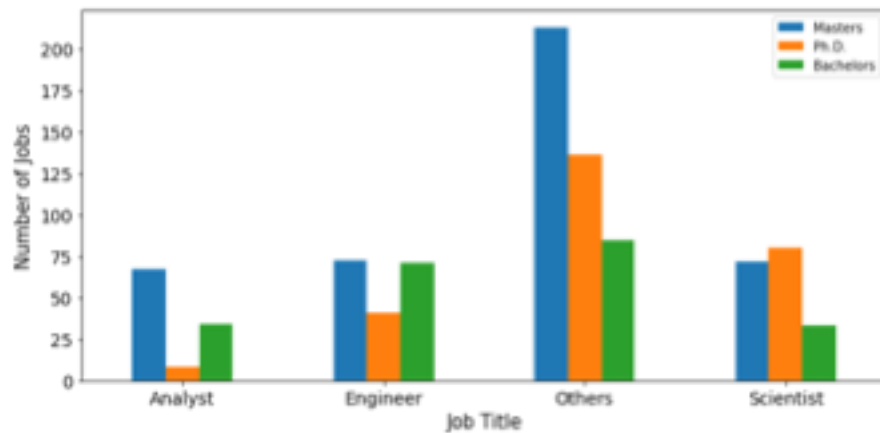
Address

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

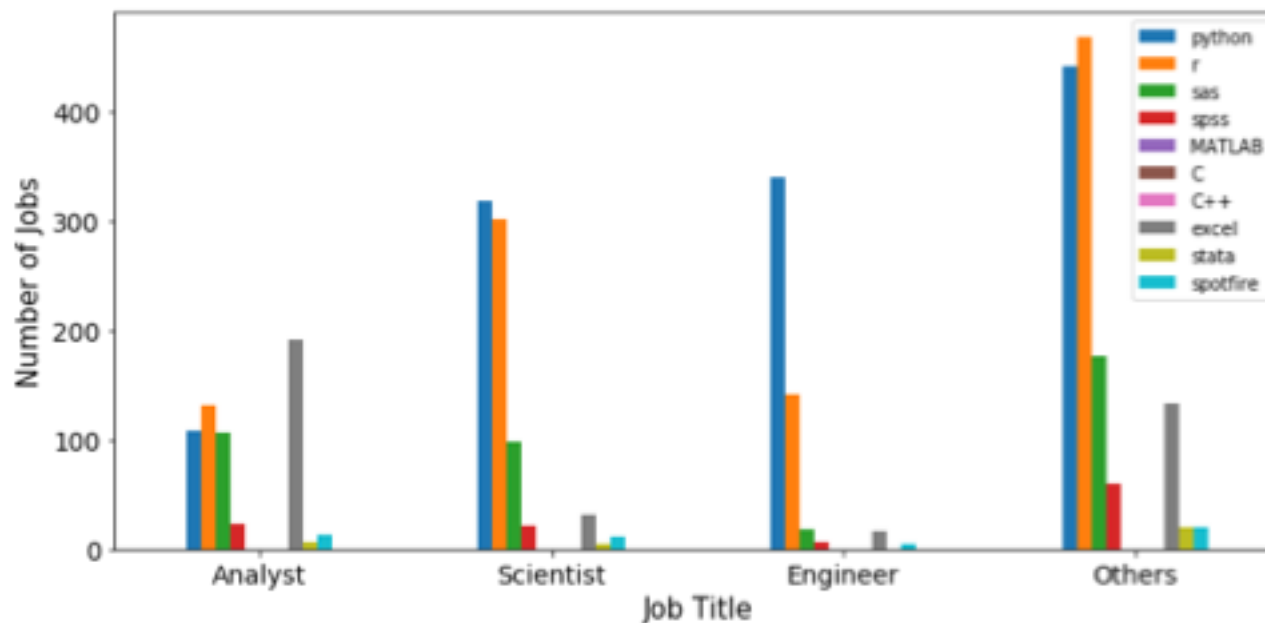
Which industry sectors are hiring data scientists?



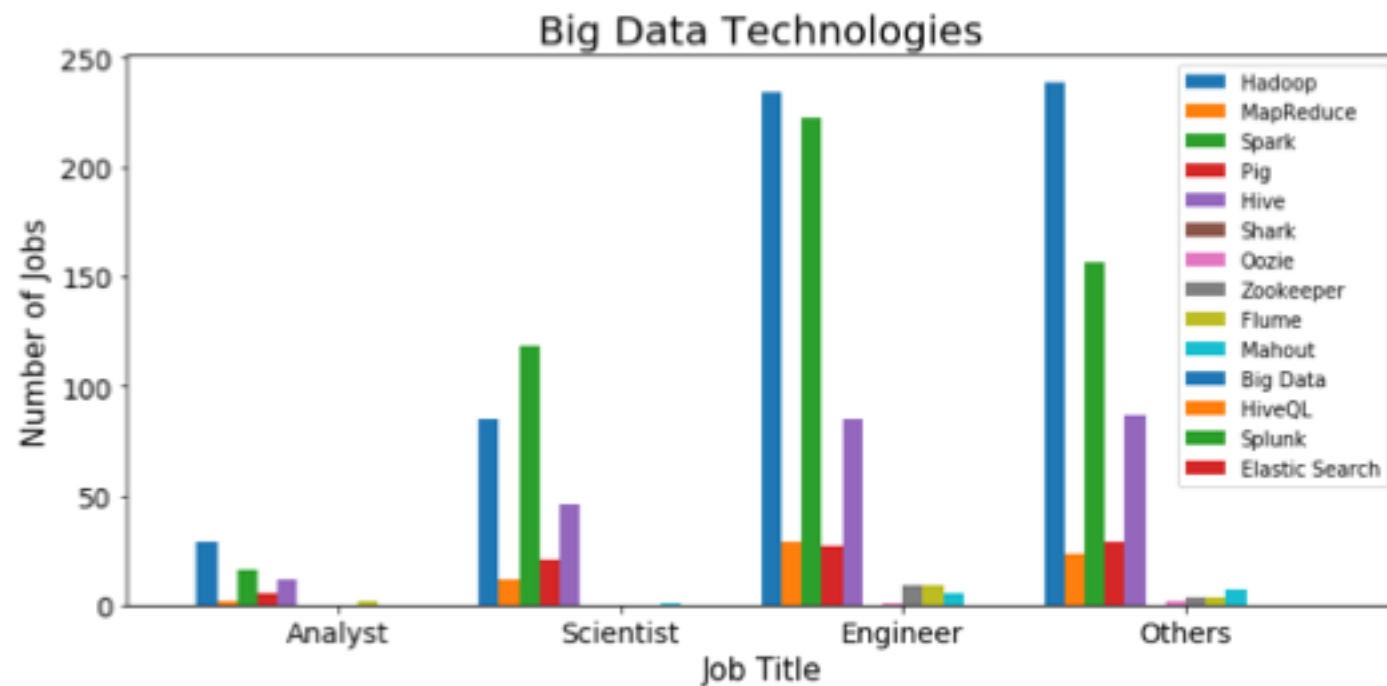
What are the educational requirements for data science jobs?



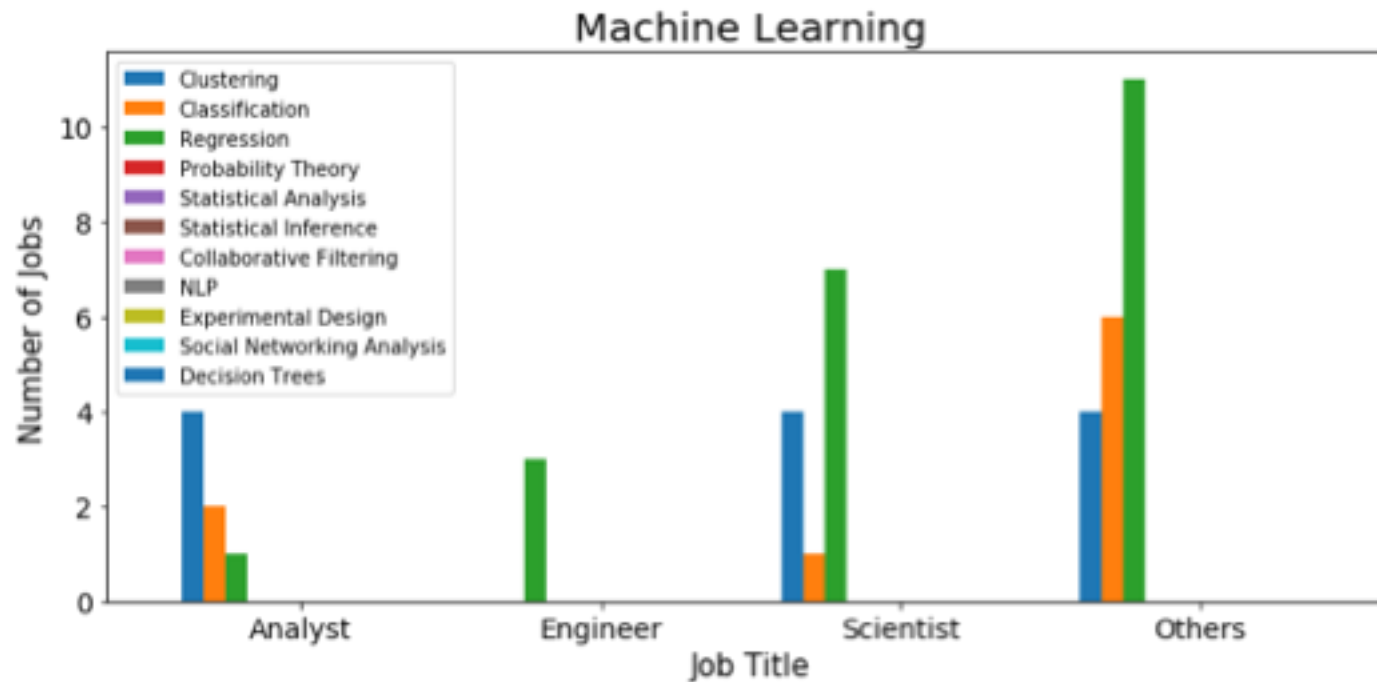
What are the important language skills for the data science jobs?



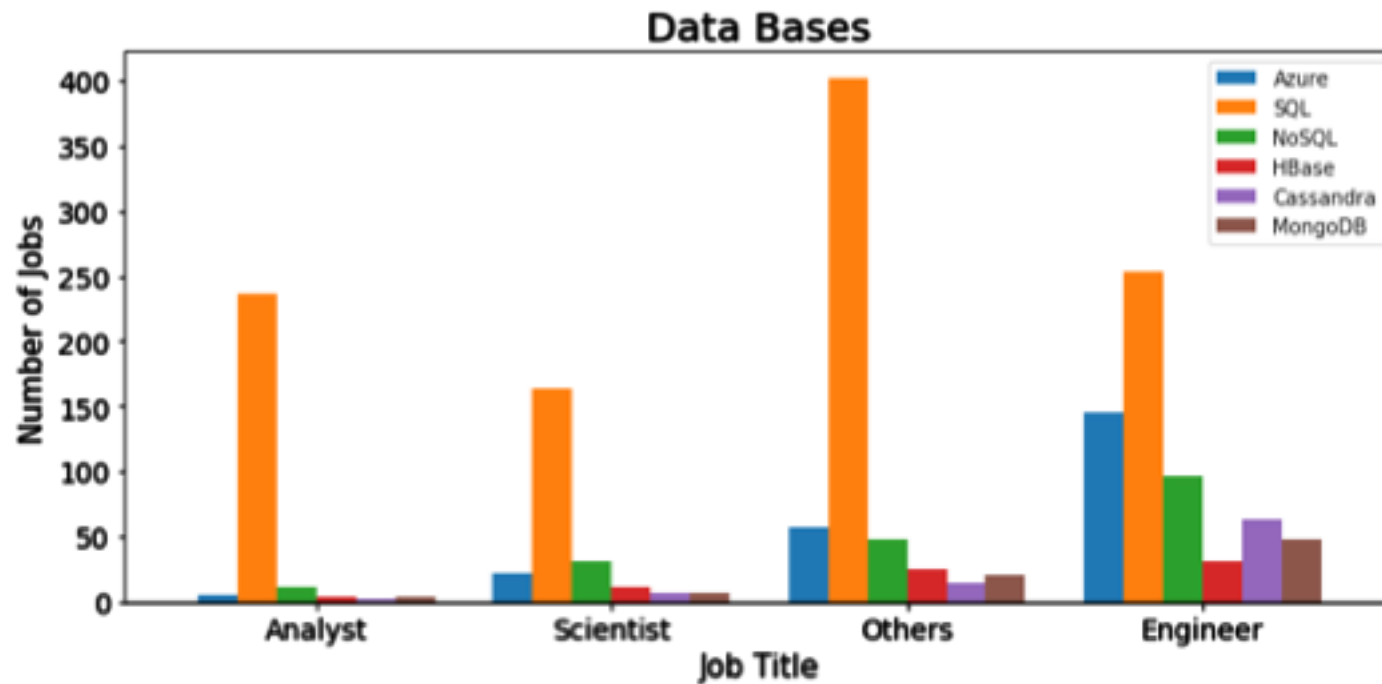
What are the important skills for the data science jobs?



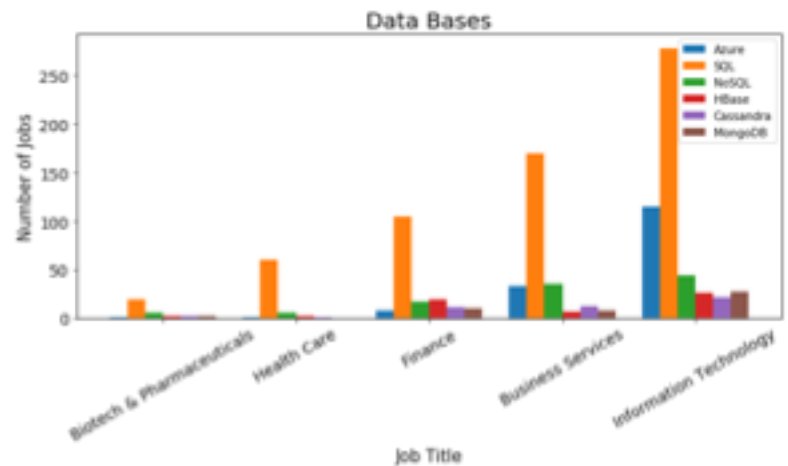
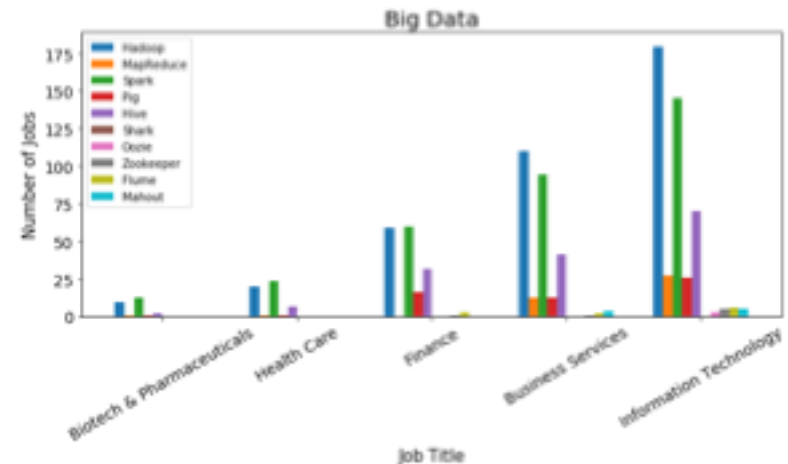
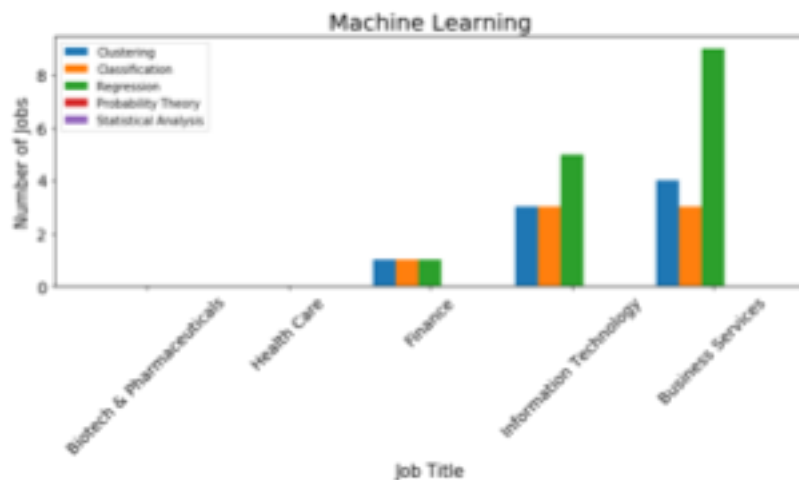
What are the important skills for the data science jobs?



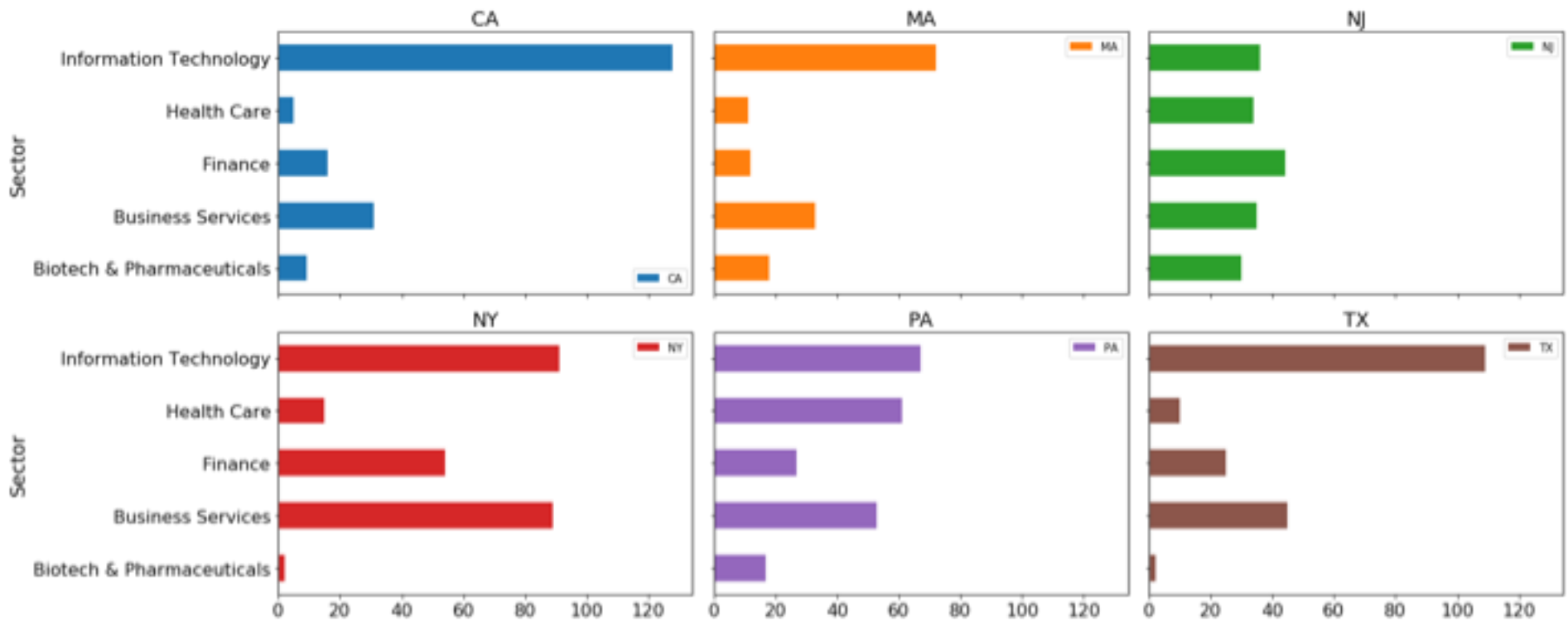
What are the important skills for the data science jobs?



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

