LinkedIn Job description analysis: Discovering Data Science Job Market Trends

- Aishwarya Budhkar
- Jagriti Kumari

Motivation

- Dynamic and Competitive nature of current data science market
- Number of data science job postings and even more number of job seekers interested in the job roles
- Teaching industry-relevant skills.
- Skillsets required for different job

Data sources

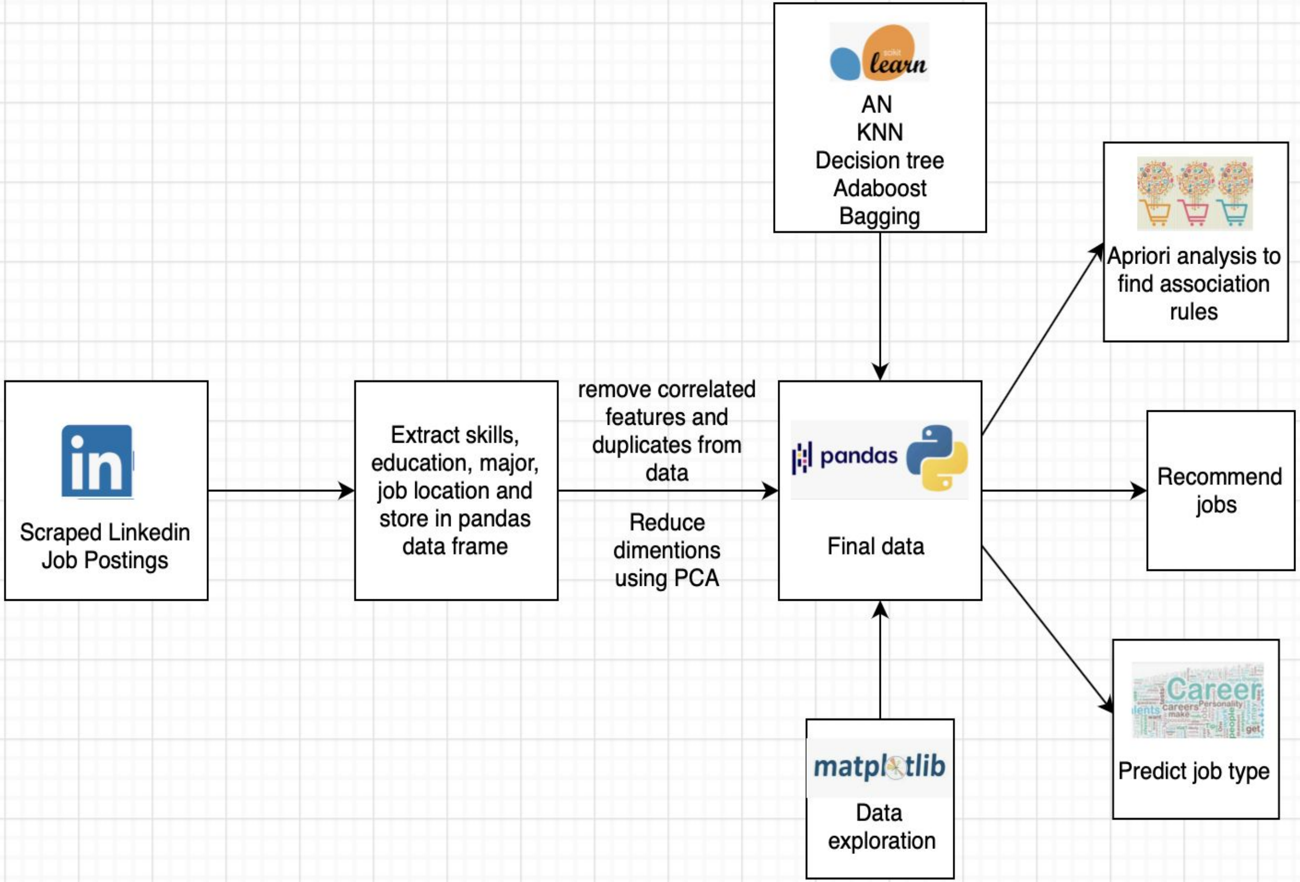
Webscrapped the data using the library linkedin-jobs-scraper 1.8.4. This library has been licensed by MIT. The dataset size was of approximately 6000 rows.

The library scrapes the publicly available job advertisements on Linkedin which includes

- Job Id
- Title
- Company
- Place
- Description
- Date
- Job function
- Employment type
- industries

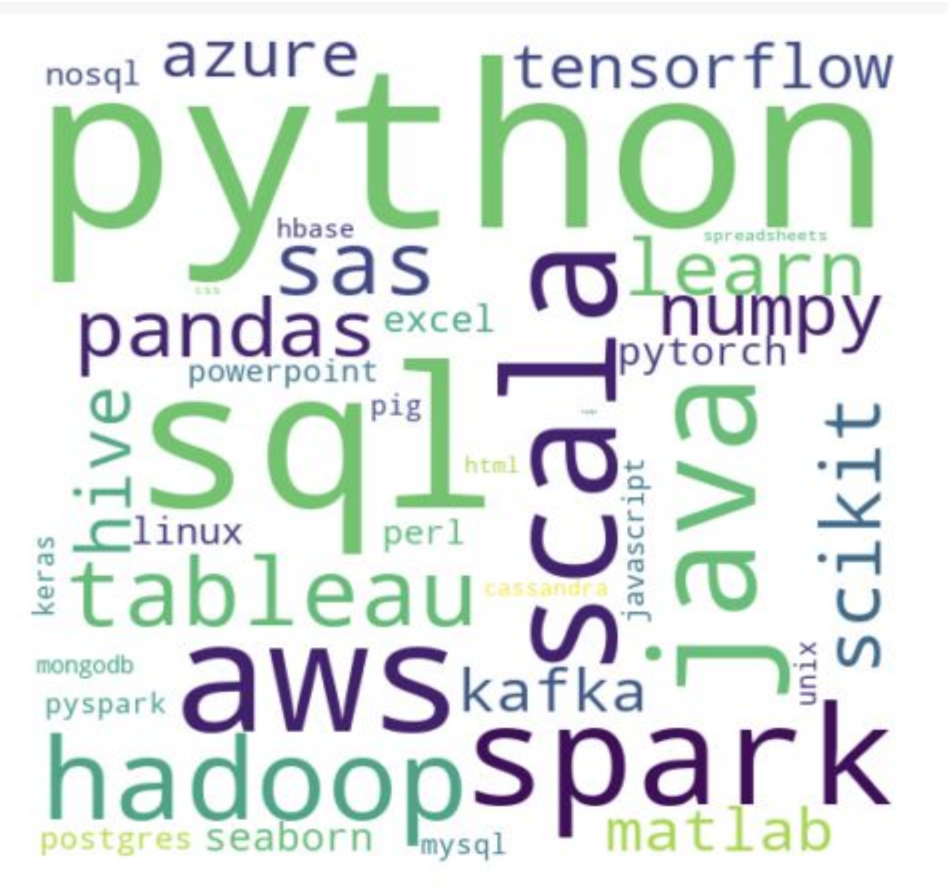
https://pypi.org/project/linkedin-jobs-scraper/

Design diagram

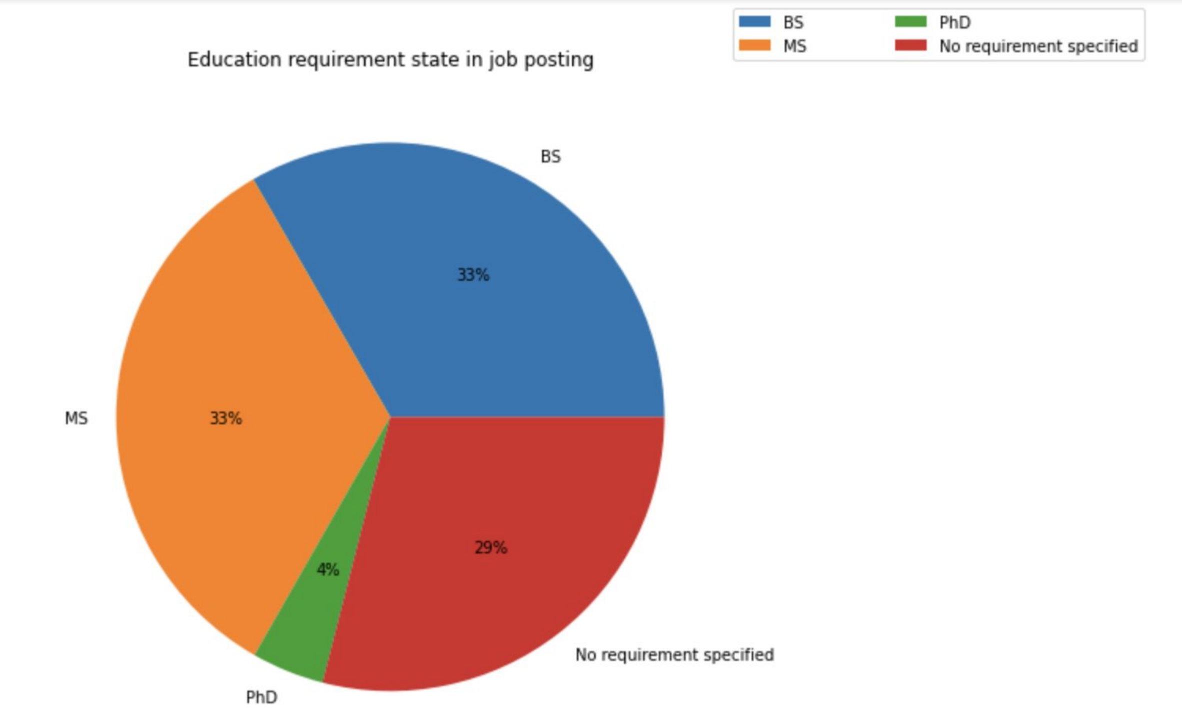


Exploratory Data analysis:

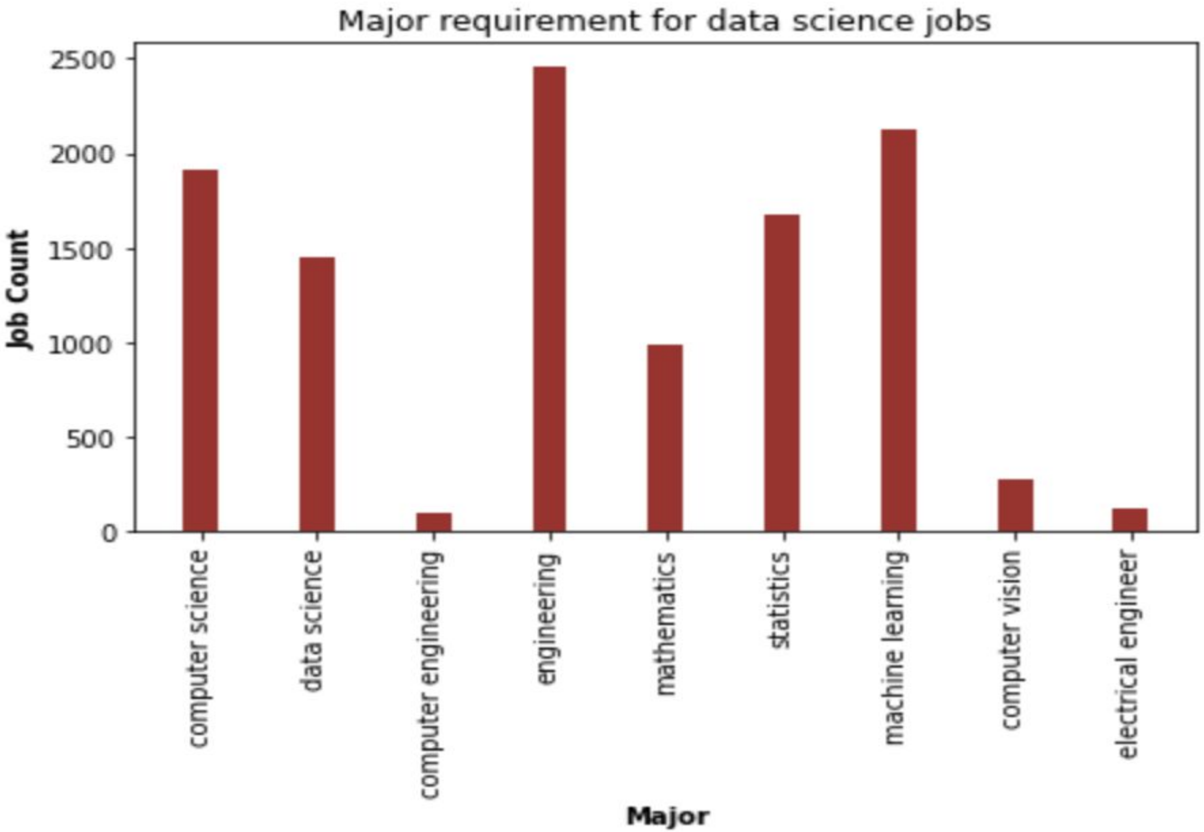
Top skills requirement



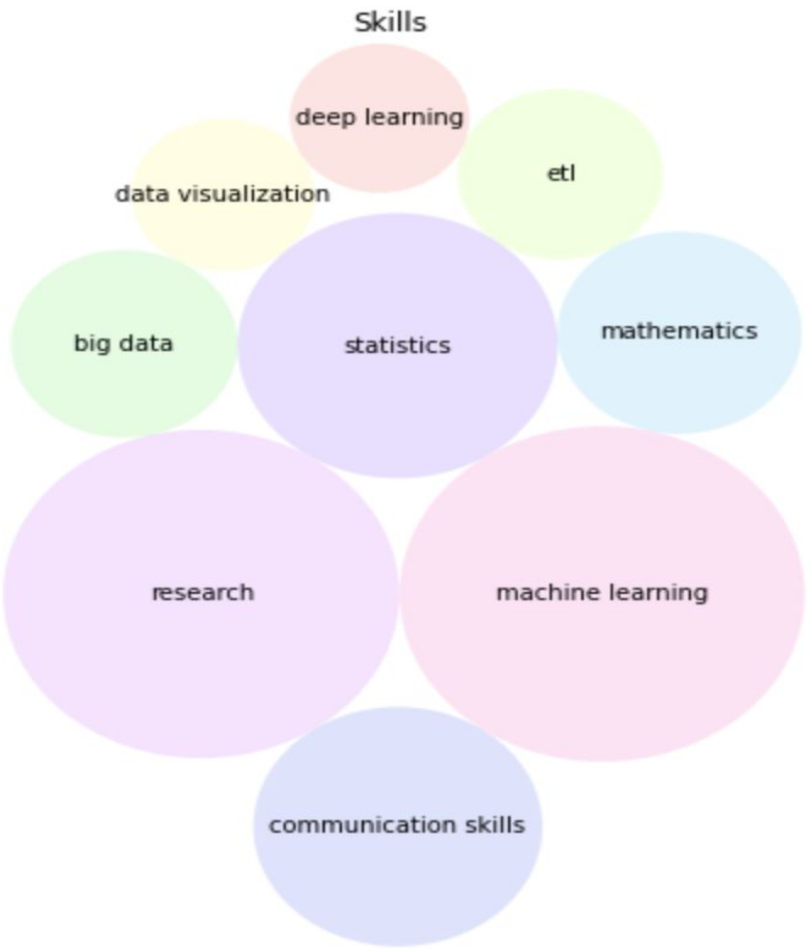
Education requirement in job postings



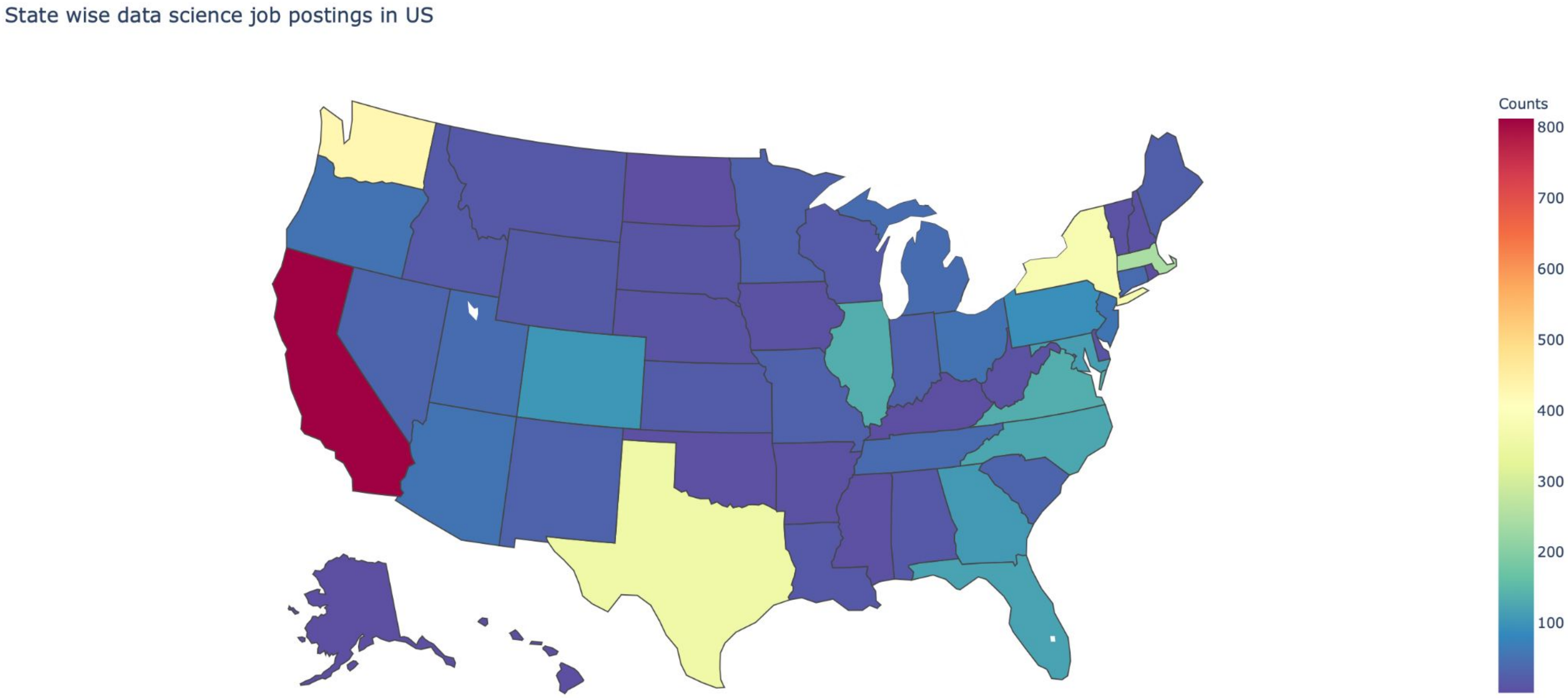
Major requirement in job postings



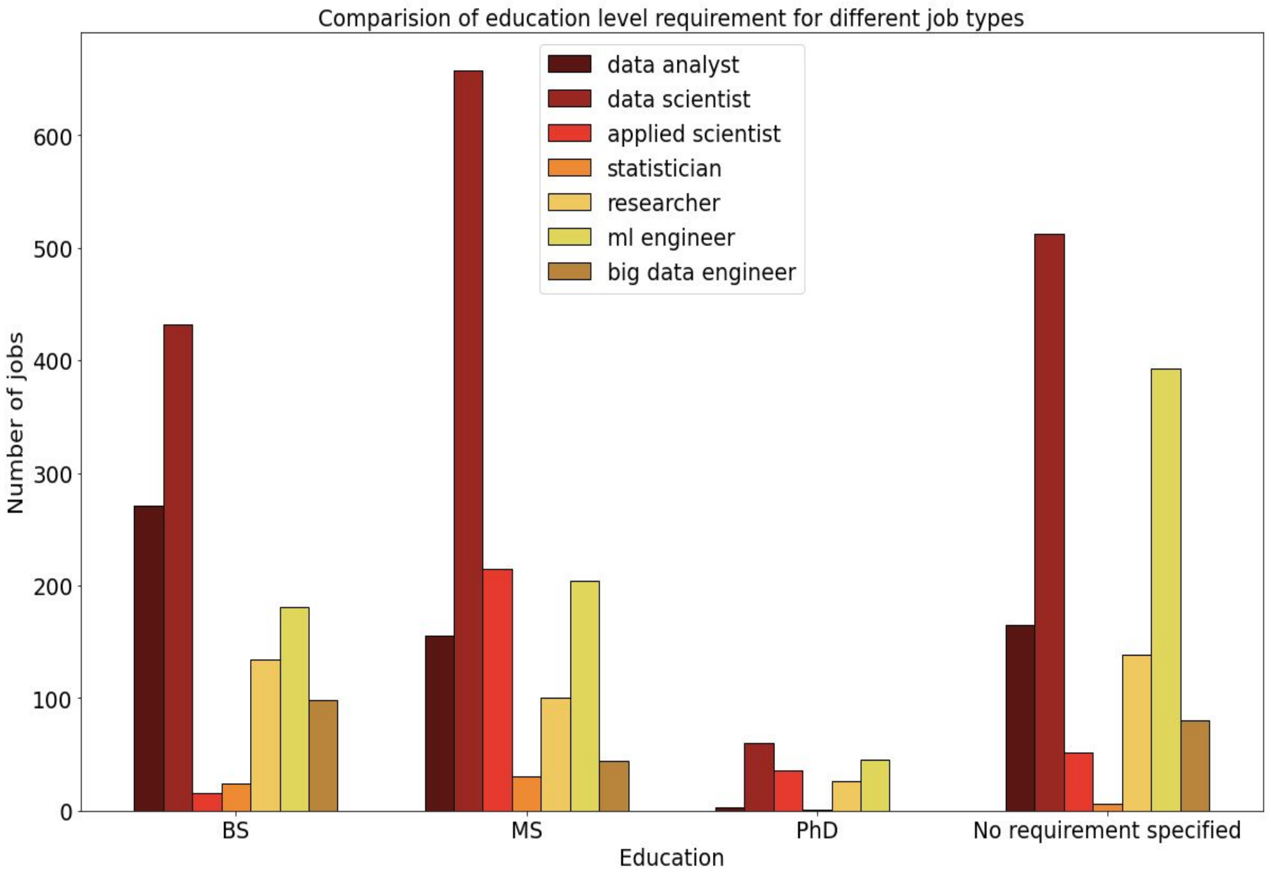
Important skills required



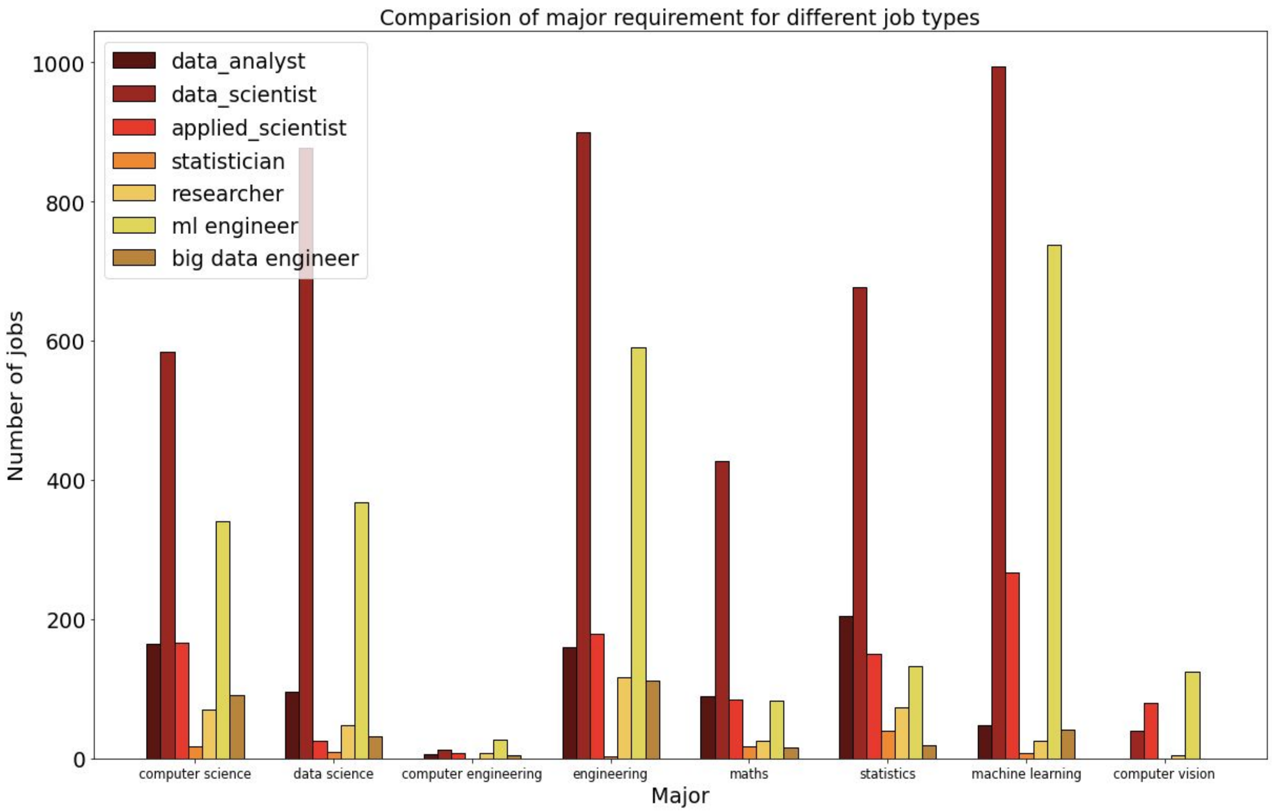
State-wise number of job-postings



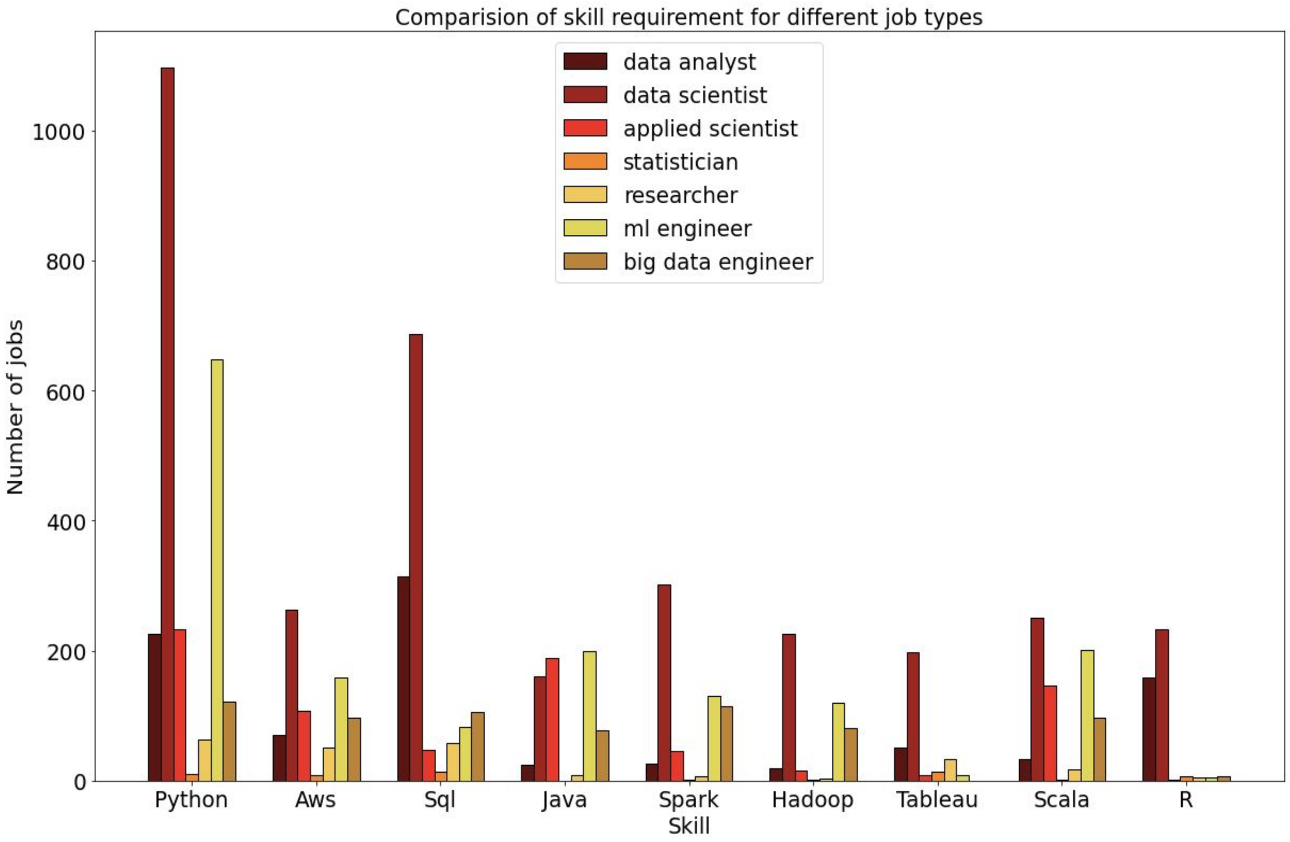
Comparison of Education requirement for different job types



Comparison of Major requirement for different job types



Comparison of Skill requirement for different job types



Insights from Data:

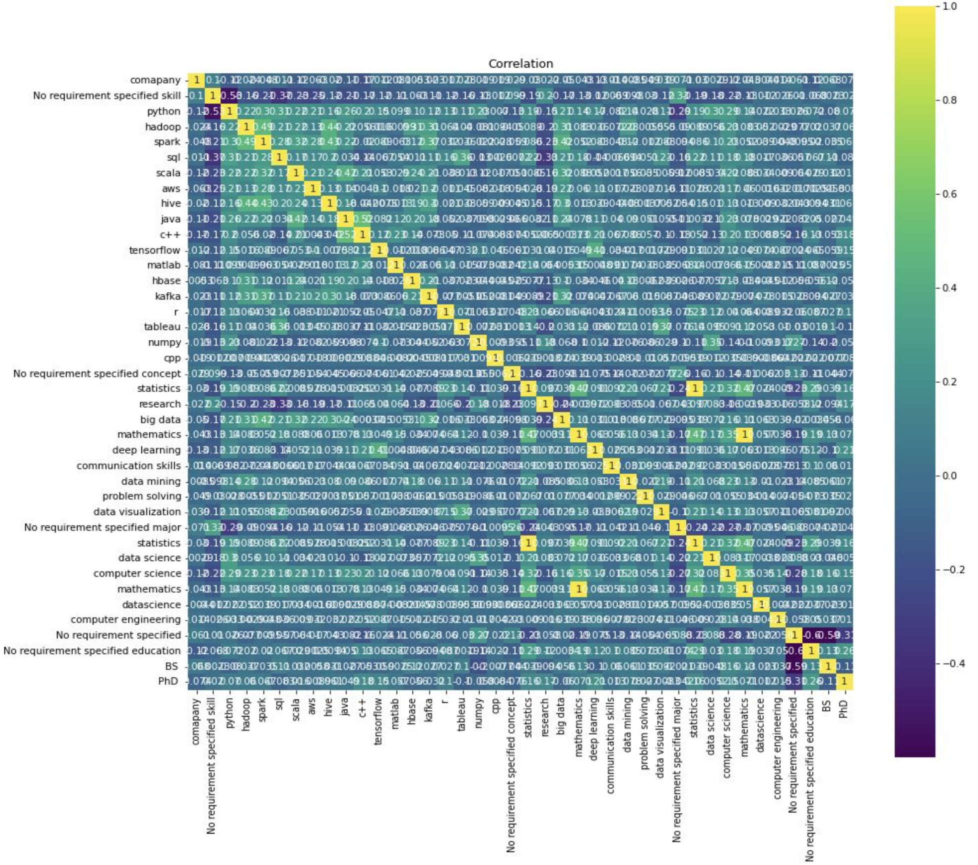
- Among technical skills, python is the most desired skill followed by Aws, Scala, Spark,Java, Hadoop and Tableau
- Bachelor and Masters graduates are equally desired. Almost ⅓ jobs do not mention degree requirement.
- Machine learning, Computer science, Data science, Statistics majors are most listed.
- Machine learning, Research, Statistics, Mathematics and communication skills are mentioned more among the jobs in data.
- Most listings are for CA followed by WA, NY, TX, IL

Insights from Data: Comparison for different job roles

- **Education**
  - Bachelor requirement for Data analyst and Big data engineer is more followed by Masters and Phd while requirement for Masters is more for Data scientist and Applied scientist.
  - Masters and Bachelors graduates are equally desired for Researcher, ML engineer and Statistician.
  - Many jobs haven't specified any education level requirement.
- **Major**
  - For ML and big data engineer machine learning and engineering majors are more important.
  - For researcher and statistician, statistics, maths and engineering majors are more desired.
  - For applied scientists and data scientists, machine learning is most important followed by data science, statistics, engineering and computer science.
  - For data analyst statistics and computer science are more desired.
- **Skills**
  - For data analyst Sql is most important skill followed by python.
  - For data scientist python is most desired followed by Sql.
  - For applied scientist, python, java, scala are most listed skills.
  - For statistician tableau is most listed followed by sql and python.
  - For researcher python,aws, sql are equally listed.
  - For ml engineer and big data engineer python, java, scala, aws,spark,hadoop are listed with python prominent is the former followed by others while all equally listed in the later.

Problem 1: Predict job type from job description

- One-hot encoded skills, education, major which lead to 102 features.
- Removed redundant and duplicate features and used data correlation to reduce features to reduce features to 50.
- Applied PCA and finally used 40 features for training the models.



Classification results

Algorithm	Test accuracy	F1 score
KNN (5 neighbors)	0.844329	0.840934
Decision tree	0.891363	0.890001
Adaboost (base estimator decision tree, 5 estimators)	**0.899896**	**0.899021**
Bagging (base estimator decision tree, 5 estimators)	0.895525	0.894282
ANN (2 hidden layers 256 128 relu activation)	85.72%	

Best performance was obtained with Adaboost with base decision tree and 5 estimators. The models were able to distinguish between the roles based on education,skills and major. Due to imbalance present in data **Synthetic Minority Over-sampling Technique** (SMOTE) was applied to balance the data for training and testing.

Problem 2: Association rule mining to find related skills

- Used **mlxtend** library.
- Generated frequent itemsets for skills in order to find the association rules between the skills.
- The frequent itemsets were one hot encoded.
- Minimum support =0.2 and minimum threshold = 0.4, got the following result:

	antecedents	consequents
0	(spark)	(python)
1	(java)	(python)
2	(sql)	(python)
3	(python)	(sql)
4	(scala)	(python)
5	(aws)	(python)

Problem 3: Job recommendation based on cosine similarity

- Implemented content based recommendation using cosine similarity.
- Data processed: normalized data
- Vectorized preprocessed job descriptions and converted it into numeric vectors.
- Computed cosine similarity on numeric vectors.
- Top 10 similar jobs.

```
job_recommender('data scientist')
```

```
array(['research analyst', 'sr. data scientist', 'senior data scientist',
       'senior research analyst', 'data scientist', 'research assistant',
       'data scientist', 'data scientist', 'data scientist'], dtype=object)
```

Obstacles

- Data collection
  - Collecting data is not easy to come by. At first, we tried to scrape data from Linkedin using selenium. However, we were limited to scrape data for only one page that consisted of only 25 job advertisements. On further research we found a library on python community that could help us scrape data for multiple pages.
- Data cleaning was challenging as it required parsing the description and extraction of required information correctly.
  - Inconsistent data. Job titles were different. So, it needed to be converted to consistent titles for the 7 job roles.
  - Duplicated data.
  - Missing values present in data. Description did not always mention all desired features.
- Non-uniform data distribution
  - Huge difference in the number of samples for each class
- Dimensionality reduction
  - Reduced features from 102 to 40 using correlation analysis and PCA

Next steps

- By collecting more data like user reviews and other user behavioral features like job likes, comments, etc. we can develop a recommendation engine to recommend relevant job to the users.
- Also, we can recommend potential candidates to recruiters and hiring managers who have required skills and a good chance of performing well in the listed job.
- More widespread data like monthly or yearly can help us analyze trends in the job posting which can help us understand when the job market is most active in the year and accordingly tailor our recommendation to job seekers who are actively and passively looking for jobs.

References

- S. Gottipati, K. J. Shim and S. Sahoo, "Glassdoor Job Description Analytics – Analyzing Data Science Professional Roles and Skills," 2021 IEEE Global Engineering Education Conference (EDUCON), 2021, pp. 1329-1336, doi: 10.1109/EDUCON46332.2021.9453931.
- R. B. Mbah, M. Rege and B. Misra, "Discovering Job Market Trends with Text Analytics," 2017 International Conference on Information Technology (ICIT), 2017, pp. 137-142, doi: 10.1109/ICIT.2017.29.
- M. Papoutsoglou, N. Mittas and L. Angelis, "Mining People Analytics from StackOverflow Job Advertisements," 2017 43rd Euromicro Conference on Software Engineering and Advanced Applications (SEAA), 2017, pp. 108-115, doi: 10.1109/SEAA.2017.50.
- Baoxu Shi, Jaewon Yang, Feng Guo, and Qi He. 2020. Salience and Market-aware Skill Extraction for Job Targeting. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '20). Association for Computing Machinery, New York, NY, USA, 2871–2879. DOI:https://doi.org/10.1145/3394486.3403338
- Katy Börner, Olga Scrivner, Mike Gallant, Shutian Ma, Xiaozhong Liu, Keith Chewning, Lingfei Wu, James A. Evans, “Skill discrepancies between research, education, and jobs reveal the critical need to supply soft skills for the data economy”, Proceedings of the National Academy of Sciences Dec 2018, 115 (50) 12630-12637; DOI:10.1073/pnas.1804247115

References

- D. Mhamdi, R. Moulouki, M.Y. El Ghoumari, M. Azzouazi, L. Moussaid, “Job Recommendation based on Job Profile Clustering and Job Seeker Behavior”, Procedia Computer Science, Volume 175,2020, Pages 695-699, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2020.07.102.
- N. A. Ismail and W. Z. Abidin, “Data Scientist Skills,” IOSR Journal of Mobile Computing & Application (IOSR-JMCA), Volume 3, Issue 4 (Jul. – Aug. 2016), PP 52-61.
- A. Shirani, “Identifying Data Science and Analytics Competencies Based on Industry Demand,” Issues in Information Systems, Volume 17, Issue IV, pp. 137-144, 2016.
- A. MOUMEN, E. H. BOUCHAMA and Y. EL BOUZEKRI EL IDIRISSI, "Data mining techniques for employability: Systematic literature review," *2020 IEEE 2nd International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS)*, 2020, pp. 1-5, doi: 10.1109/ICECOCS50124.2020.9314555.
- I. Khaouja, I. Rahhal, M. Elouali, G. Mezzour, I. Kassou and K. M. Carley, "Analyzing the needs of the offshore sector in Morocco by mining job ads," 2018 IEEE Global Engineering Education Conference (EDUCON), 2018, pp. 1380-1388, doi: 10.1109/EDUCON.2018.8363390.

THANK YOU