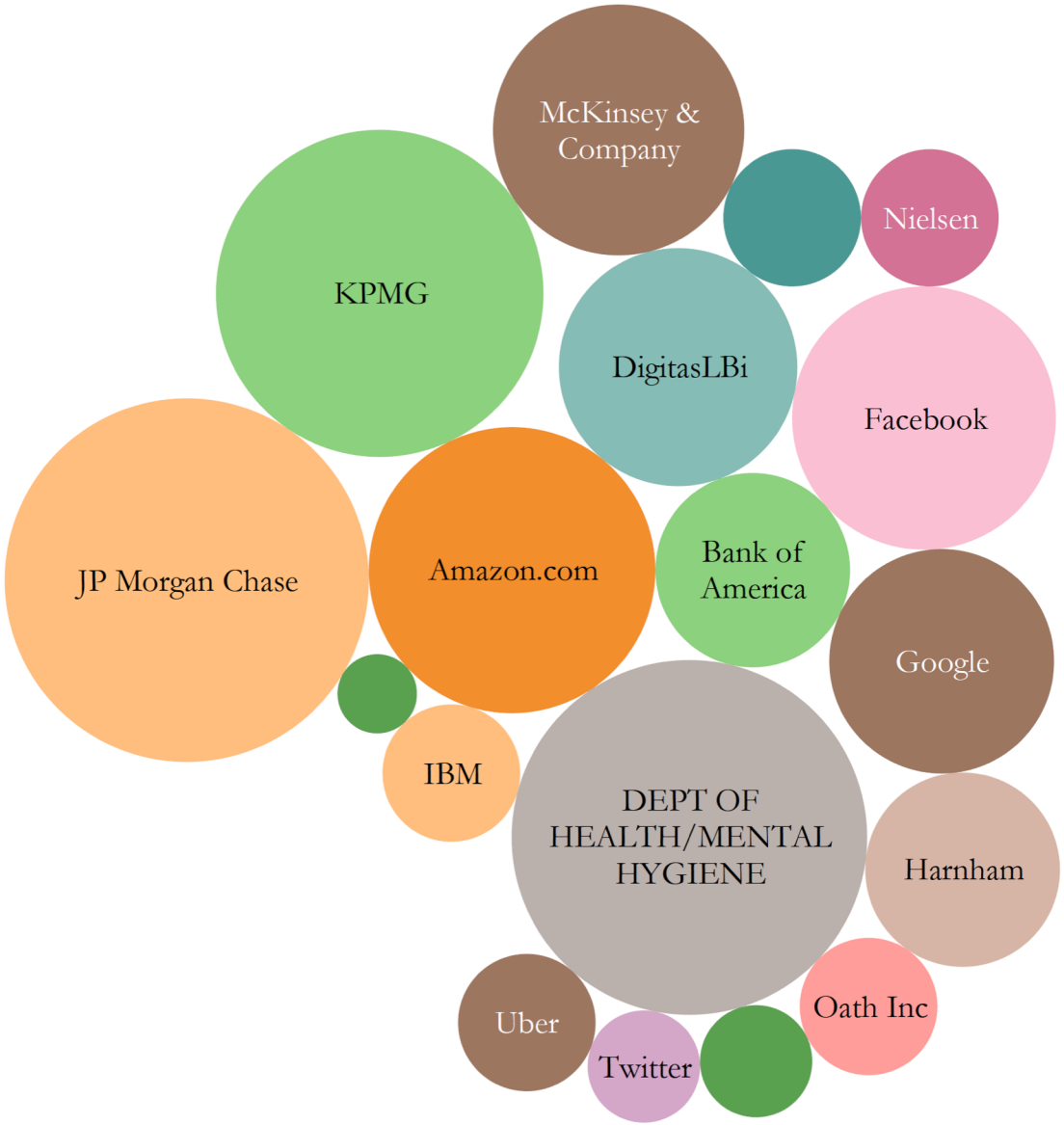
# Who gets hired?

An outlook of the Data Scientist Job Market in the U.S.

NYC Data Science Academy #14

Shanshan "Silvia" Lu

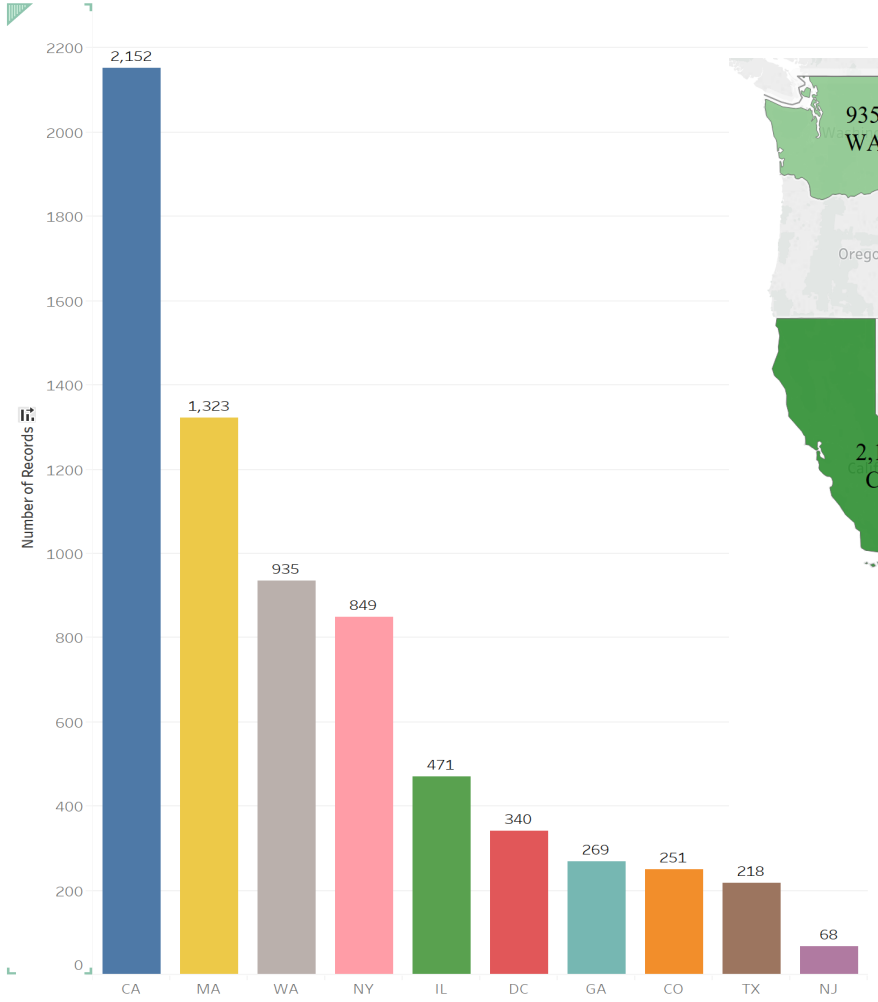


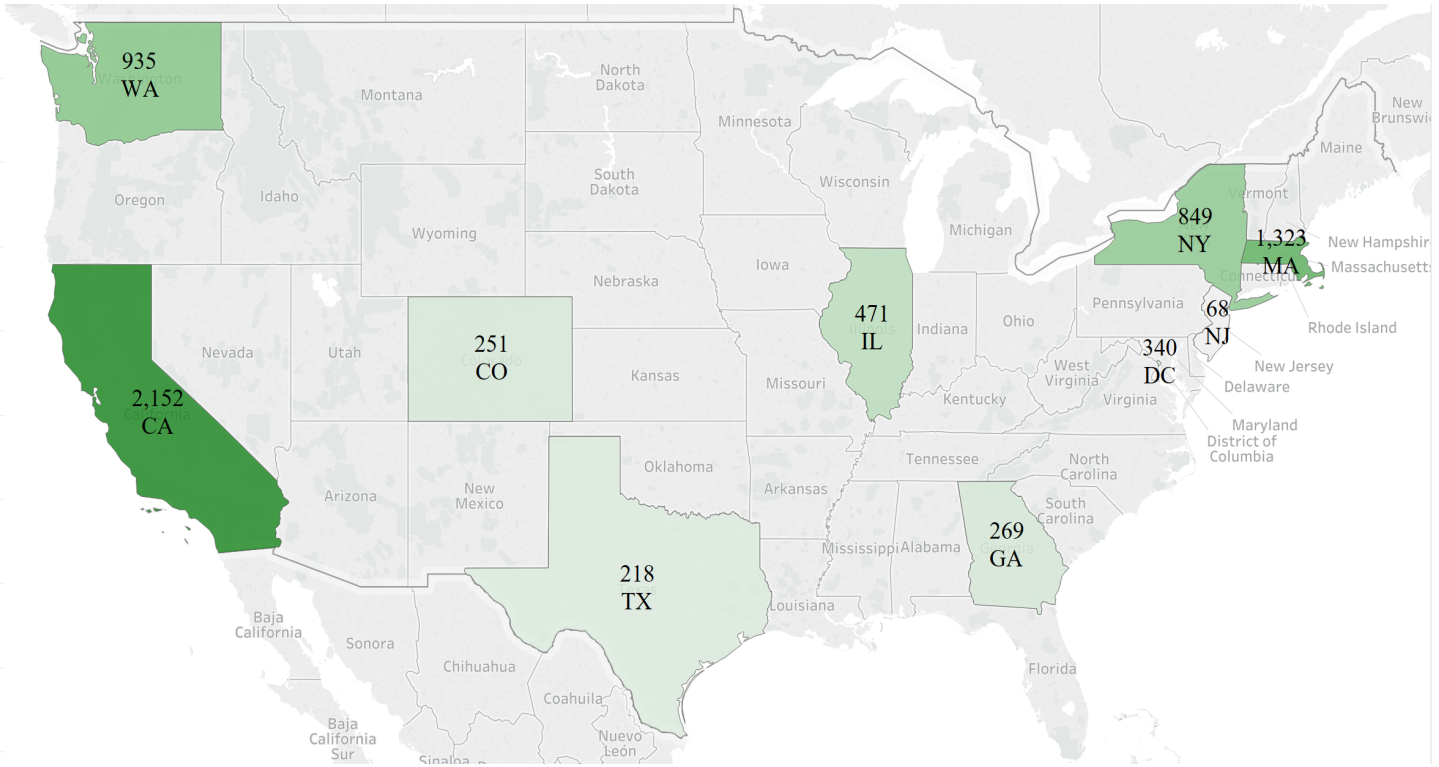
# Scraping Process and Problems

- Indeed limits the pages to 100 at most.
- Recurrent sponsored jobs.

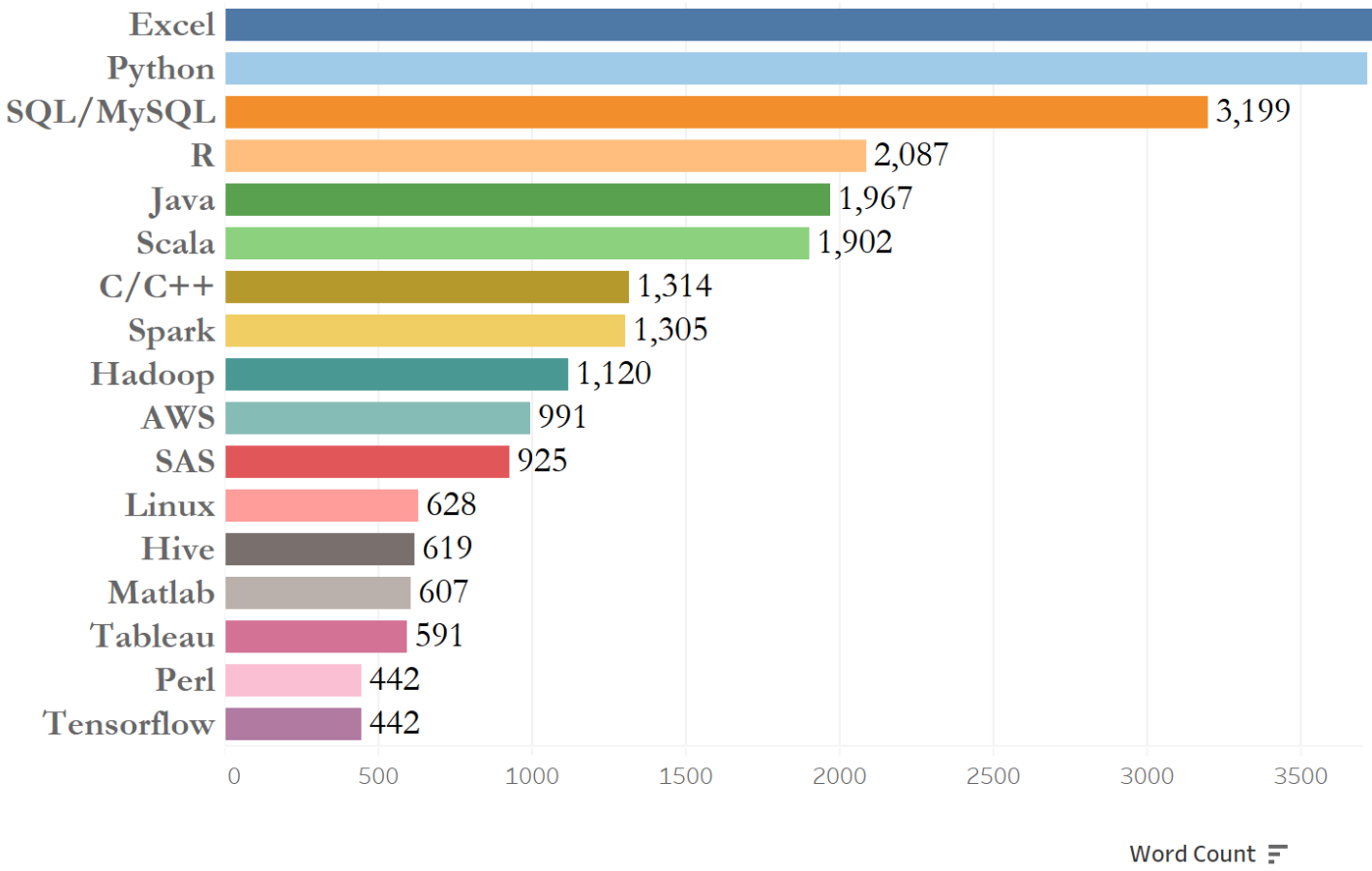
- Data Source: https://www.indeed.com/
- Scraped Information: Position, Company, Location, Job Description, and Number of Company Reviews.
- In total, **6876** jobs have been scraped on Aug. 3rd.
- Exploring questions:
  - What type of talents do employers want concerning tools, skills, degrees, and majors?
  - Do employers know the difference between data engineer, data scientist and data analyst? Does the job description match the position?

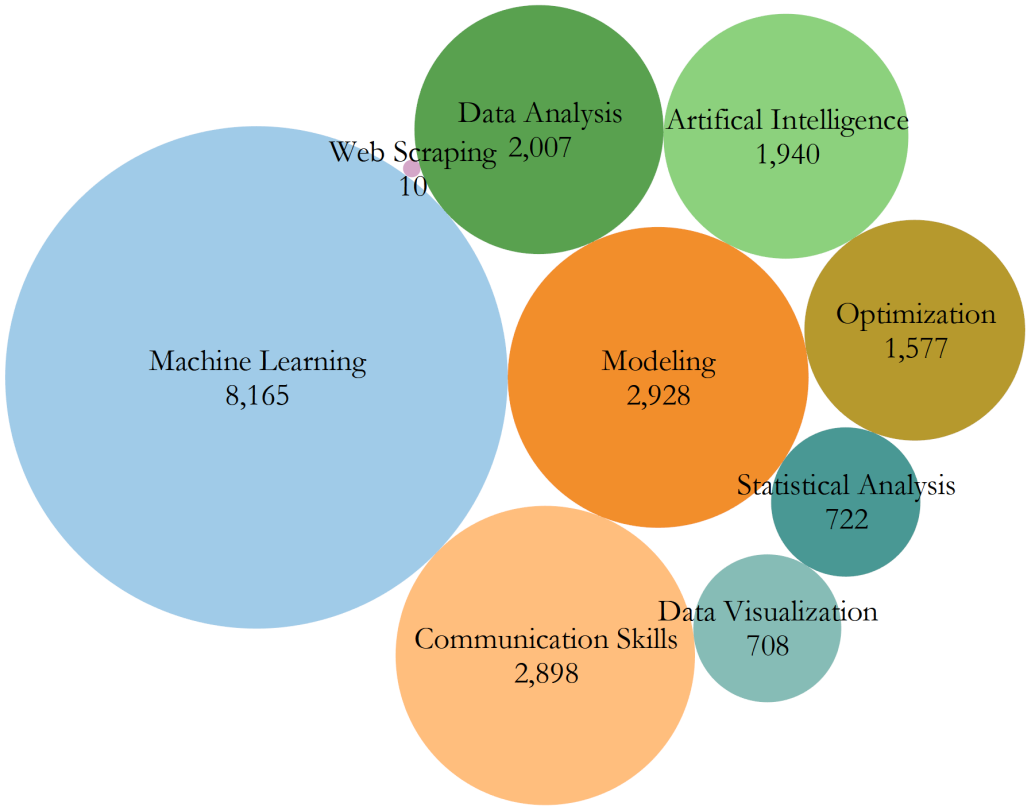
# Number of Data Scientist Job in 2018

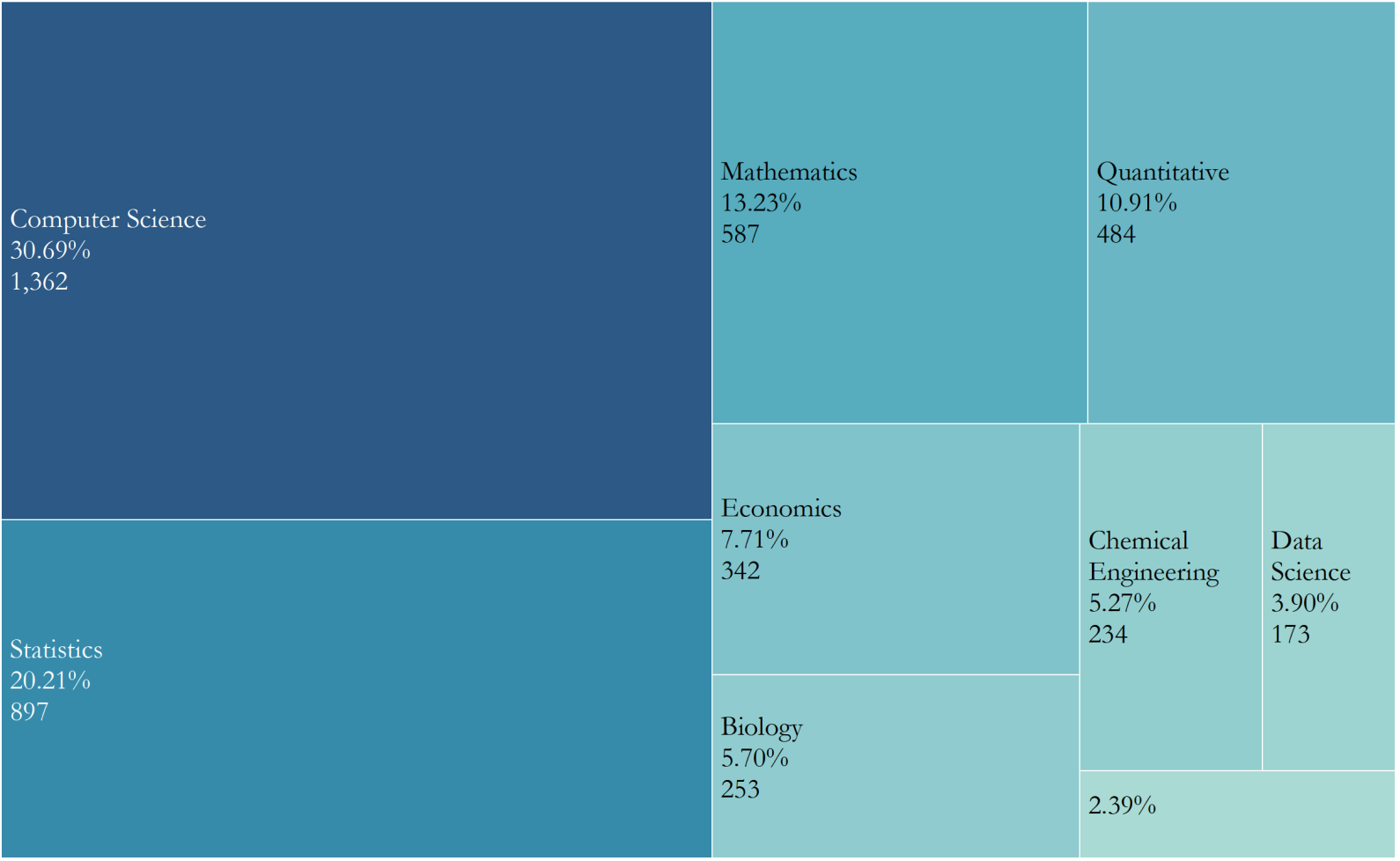


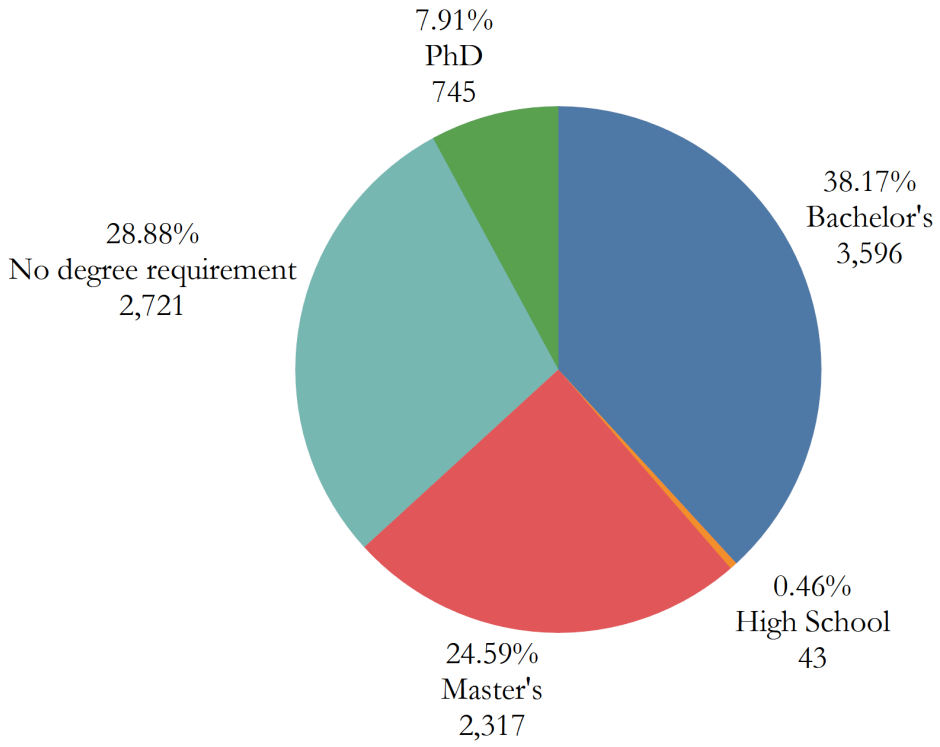
# What tools and skills are desired the most?



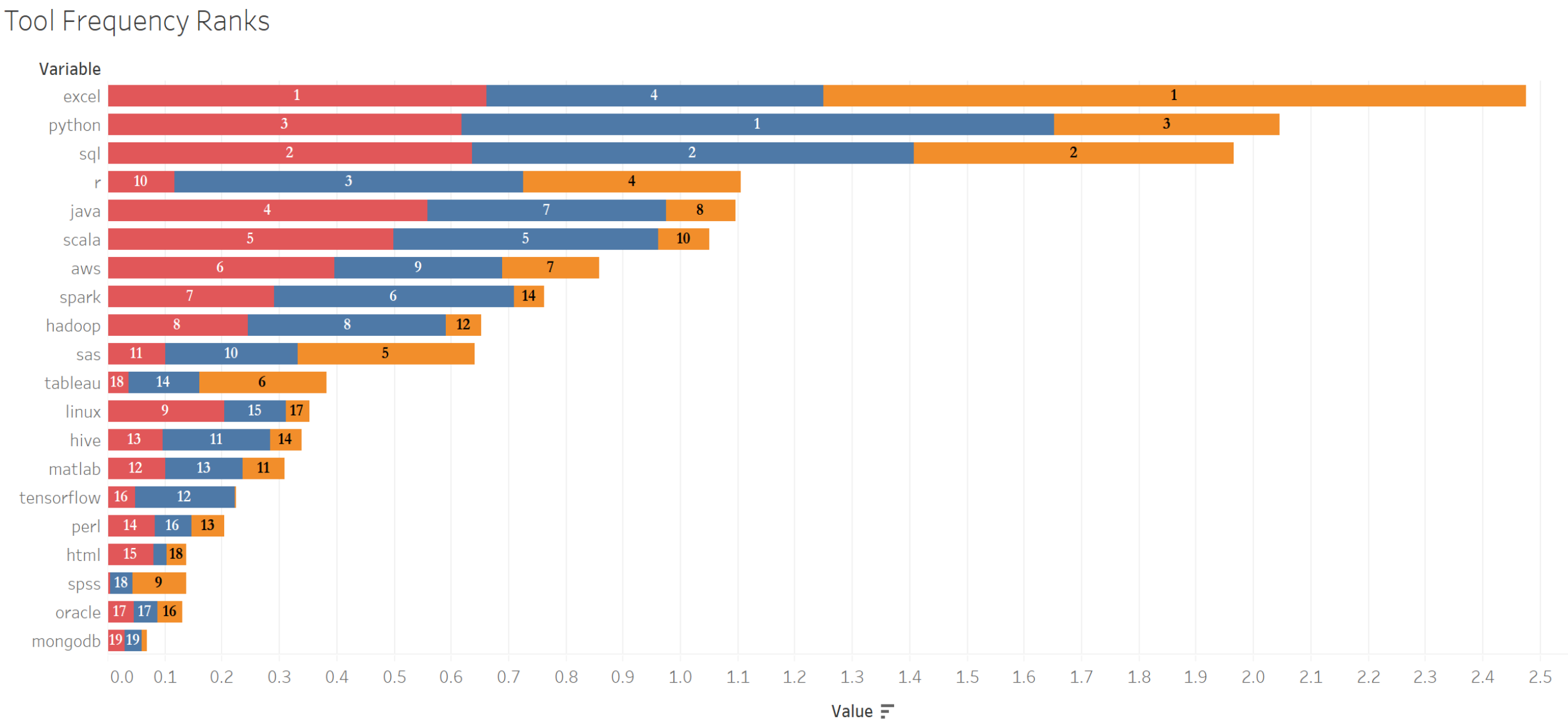






# Does the degree matter?

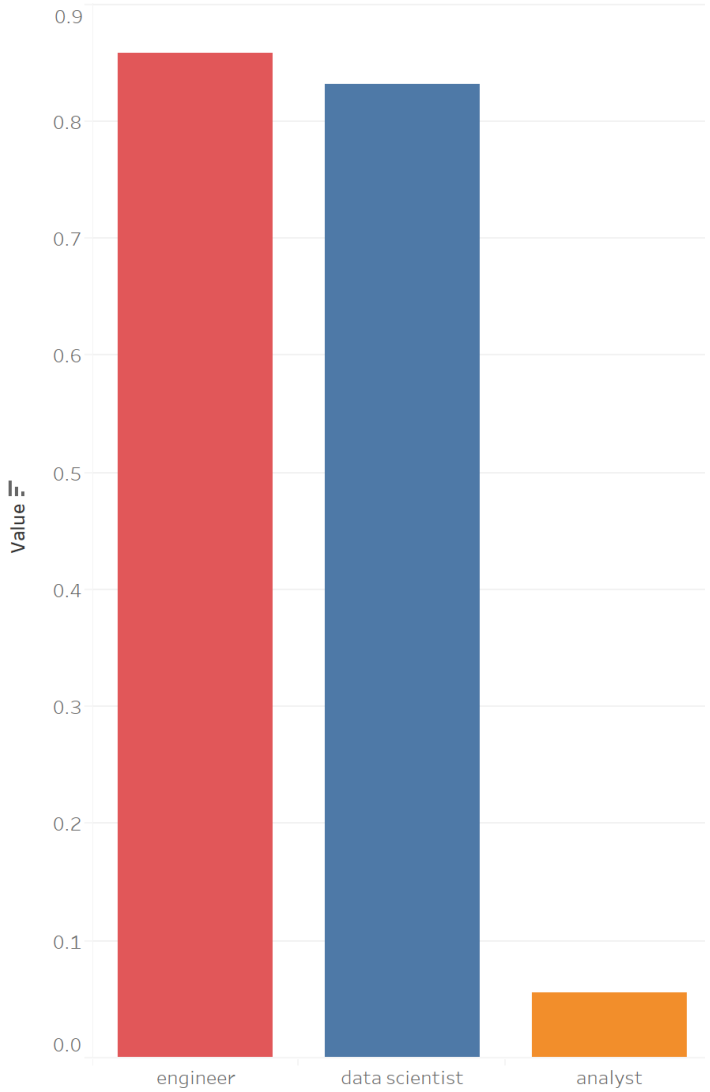


# Engineer vs. Data Scientist vs. Analyst

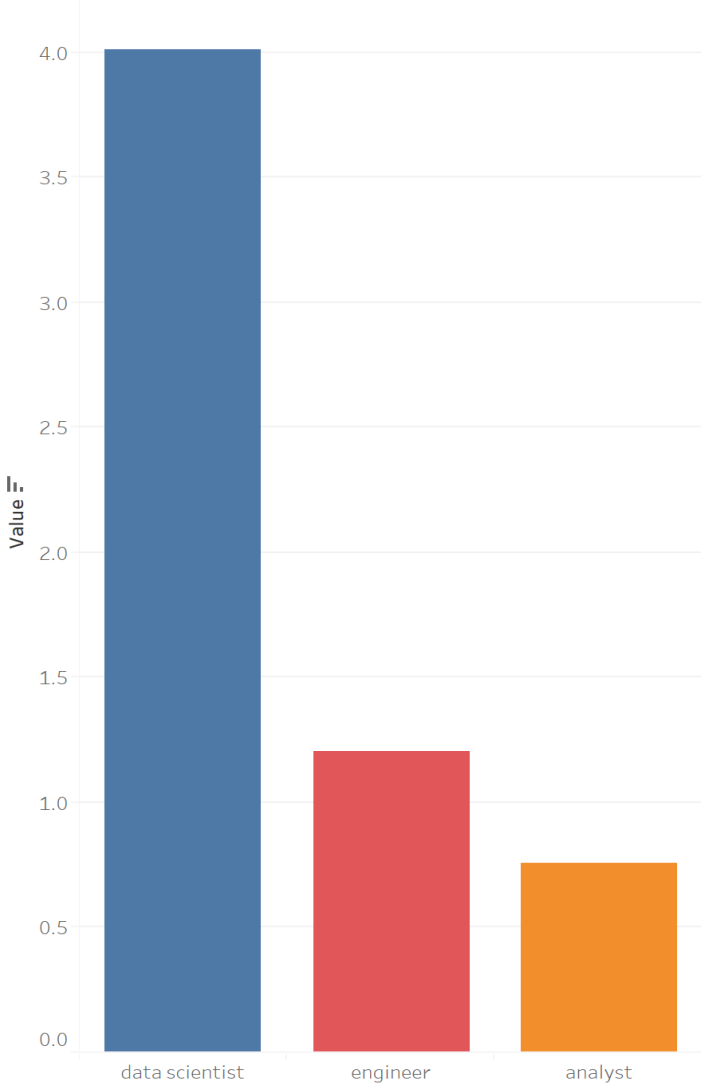


# Engineer vs. Data Scientist vs. Analyst

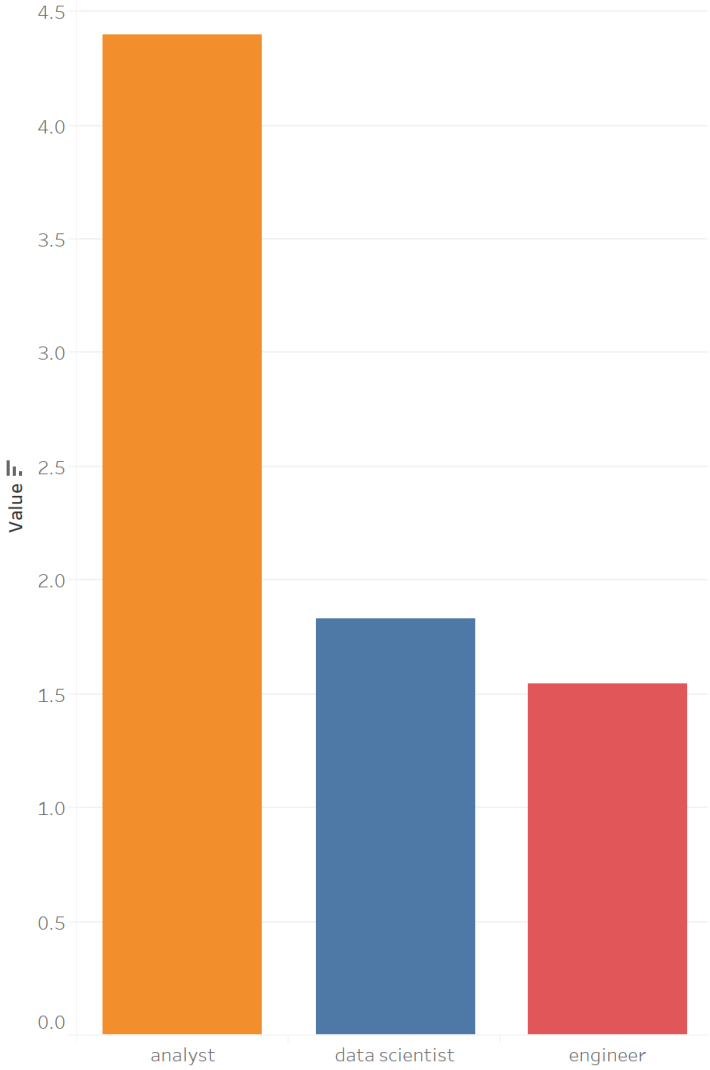
Sum of Keywords Frequency: Artificial Intelligence



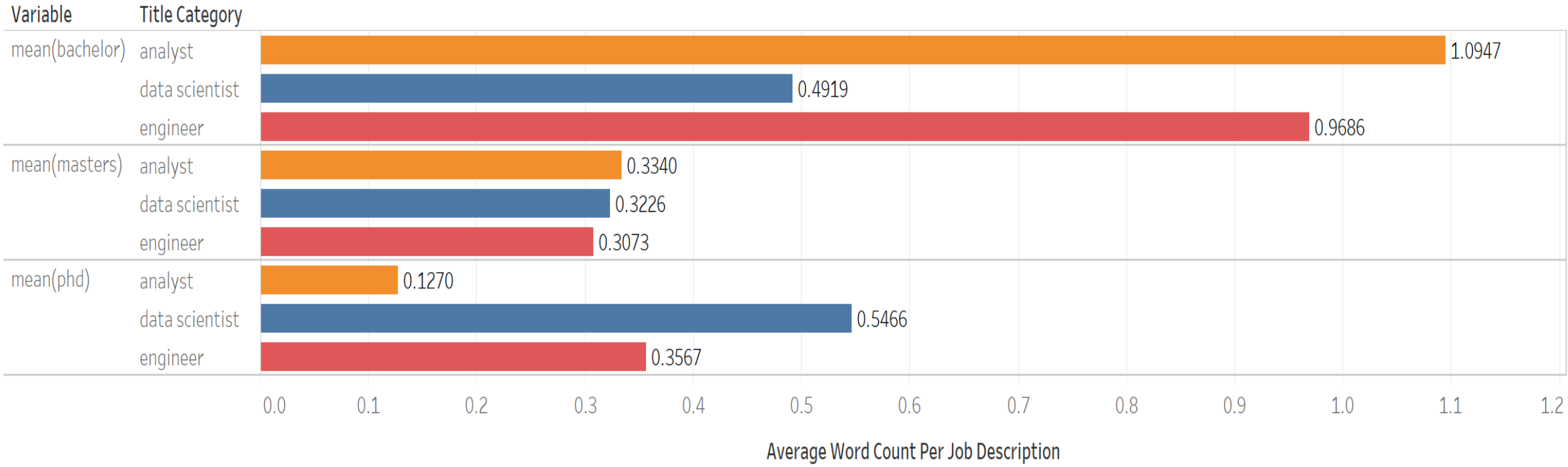

Sum of Keywords Frequency: Machine Learning, Modeling, Statistical Analysis




Sum of Keywords Frequency: Data Analysis, Data Visualization, Research.

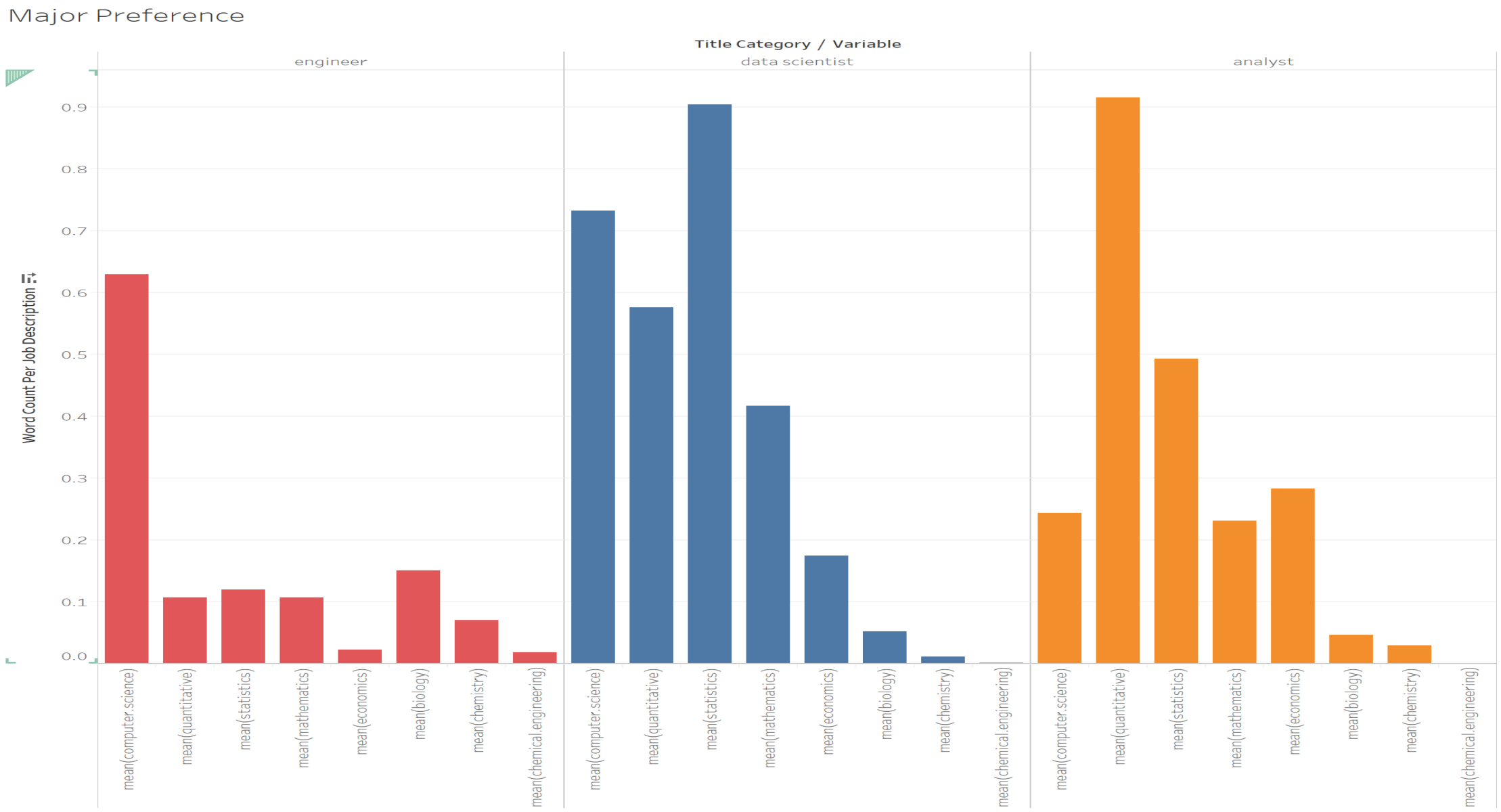


# Engineer vs. Data Scientist vs. Analyst



# Engineer vs. Data Scientist vs. Analyst

## Majors



# Can we use word frequencies to classify job types?

| python | excel | sql | java | scala | spark | machine learning | data analysis | data visualization | modeling |
|---|---|---|---|---|---|---|---|---|---|
| 3 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 2 |
| 2 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 2 | 3 |
| 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 |
| 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

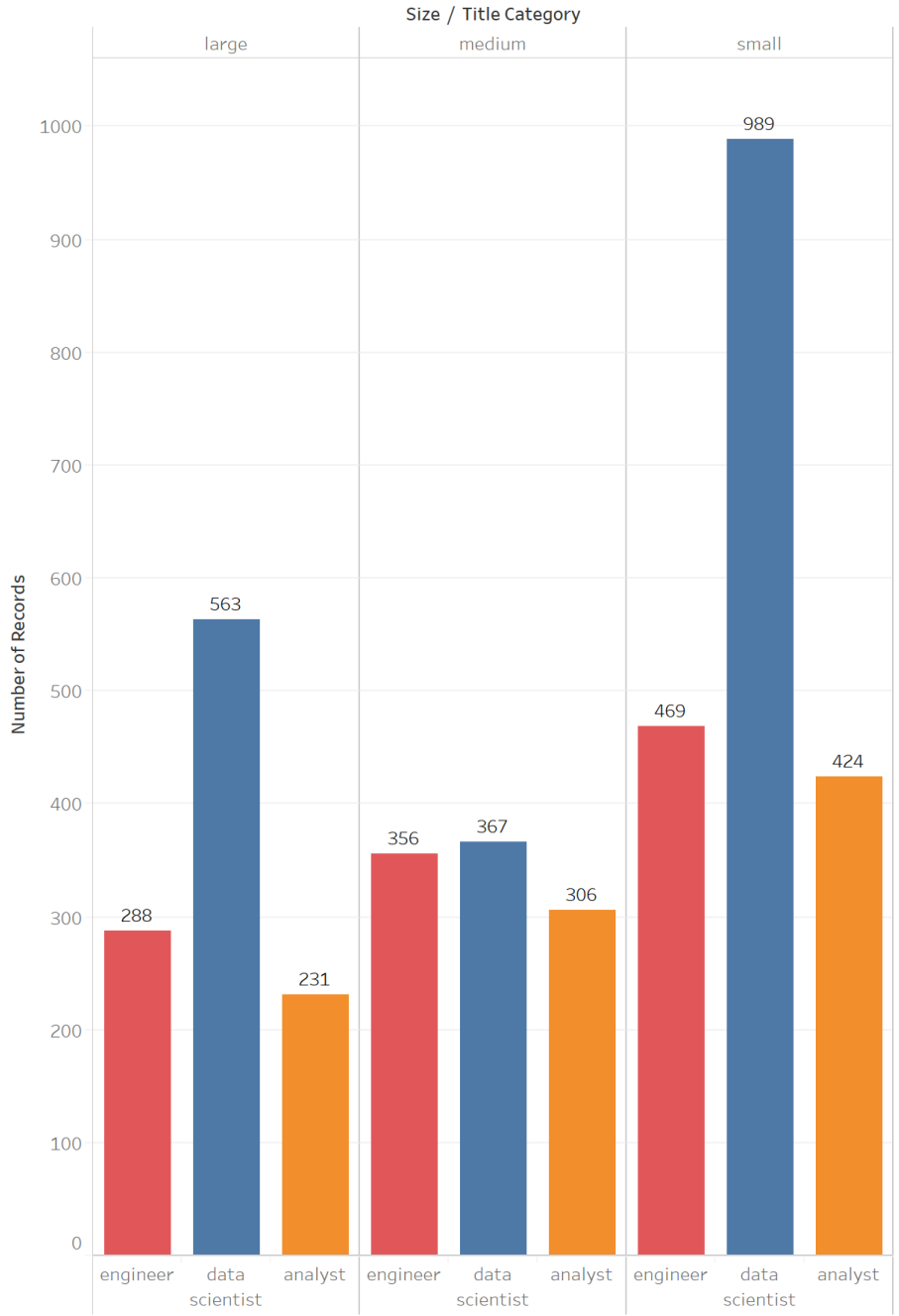
- Problem:
  - All variables are picked manually.
  - Too many variables for an unsupervised machine learning clustering.
- Results:

```
In [51]: 1 logit_1.score(indeed_X,knowndata['target'])

Out[51]: 0.7773603806661658
```

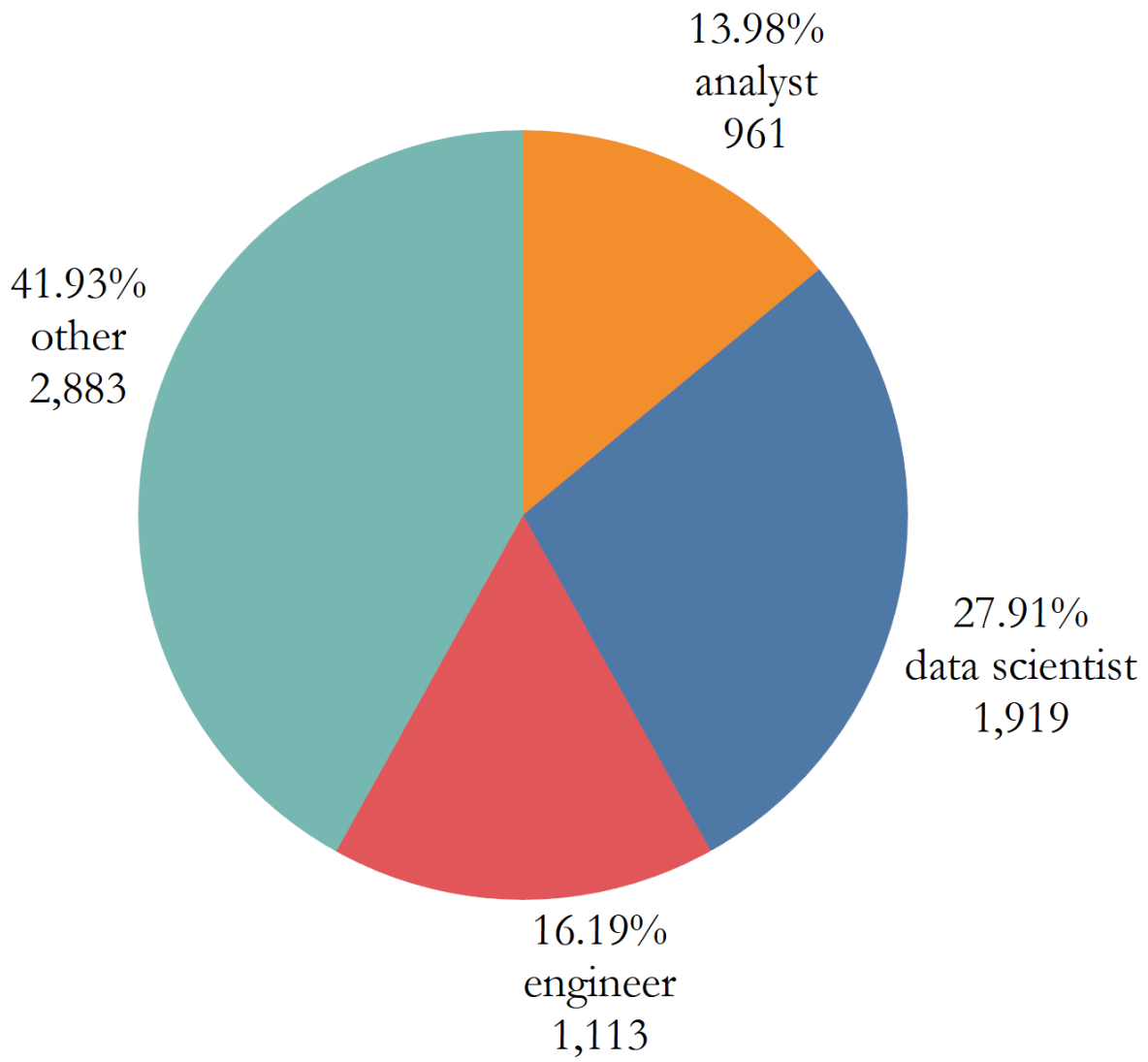
- Possible Reasons:
  - Predictors were chosen badly.
  - Employers don't know what type of people they want.

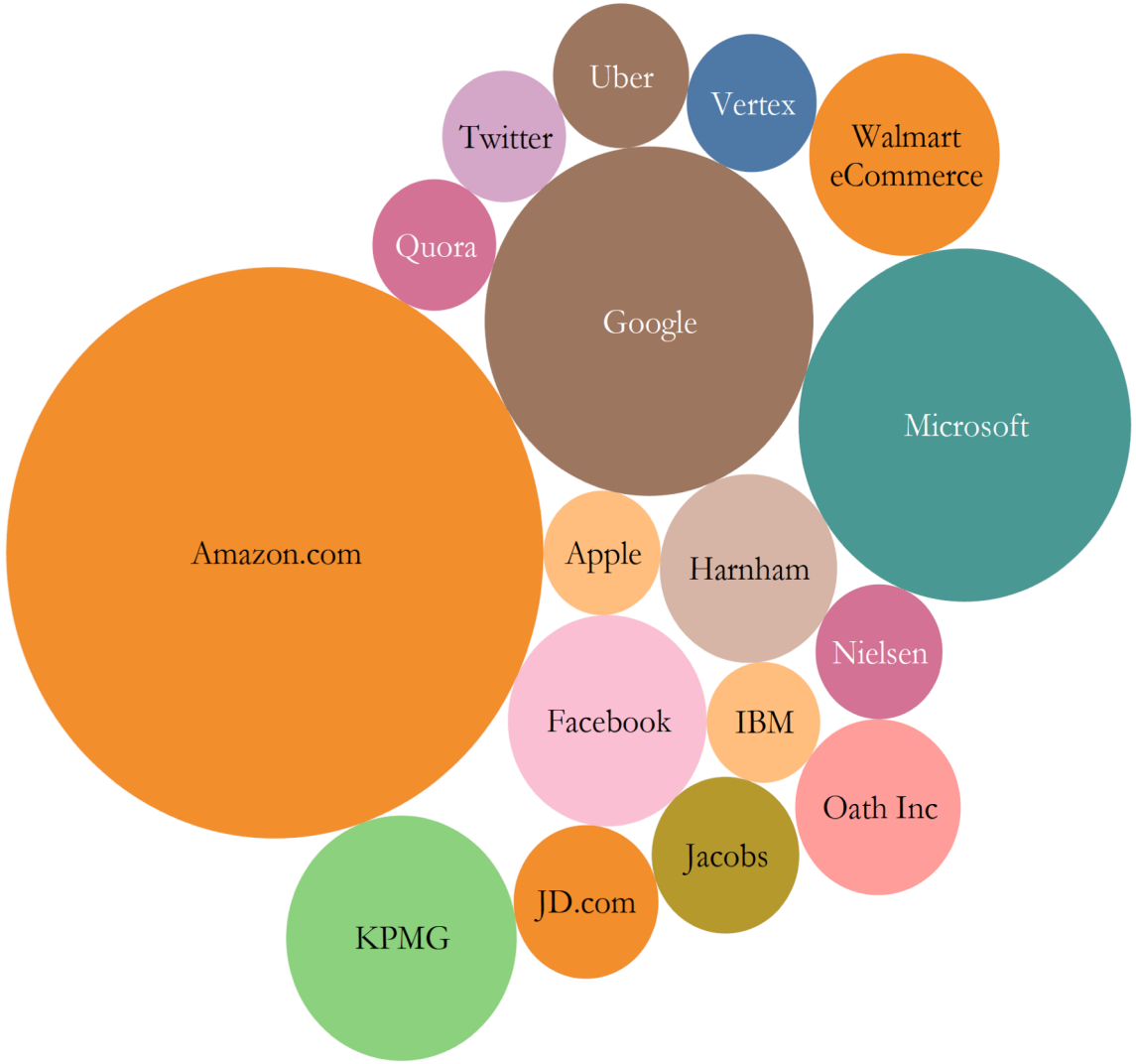
# Small companies hiring the most data scientists?



# Limitations

- Customized job postings by indeed may bias the data.
- Limited sample size.
- Unverified title categorization.
- Overcounting several words: “engineer”, “data science”, etc.






# Conclusions

- Who gets hired:
  - Excel, Python, R, SQL,
  - Machine Learning, Data/Statistical Analysis, Data Visualization.
  - Master's degree (or higher) in Statistics, Mathematics, Computer Science, Data science, Engineer or Quantitative field.
  - (Resume tips)
- Do employers know who they want to hire?
  - Not really : (
  - Read job descriptions carefully.