DATA SCIENCE SKILLS THAT MATTER

Project 3

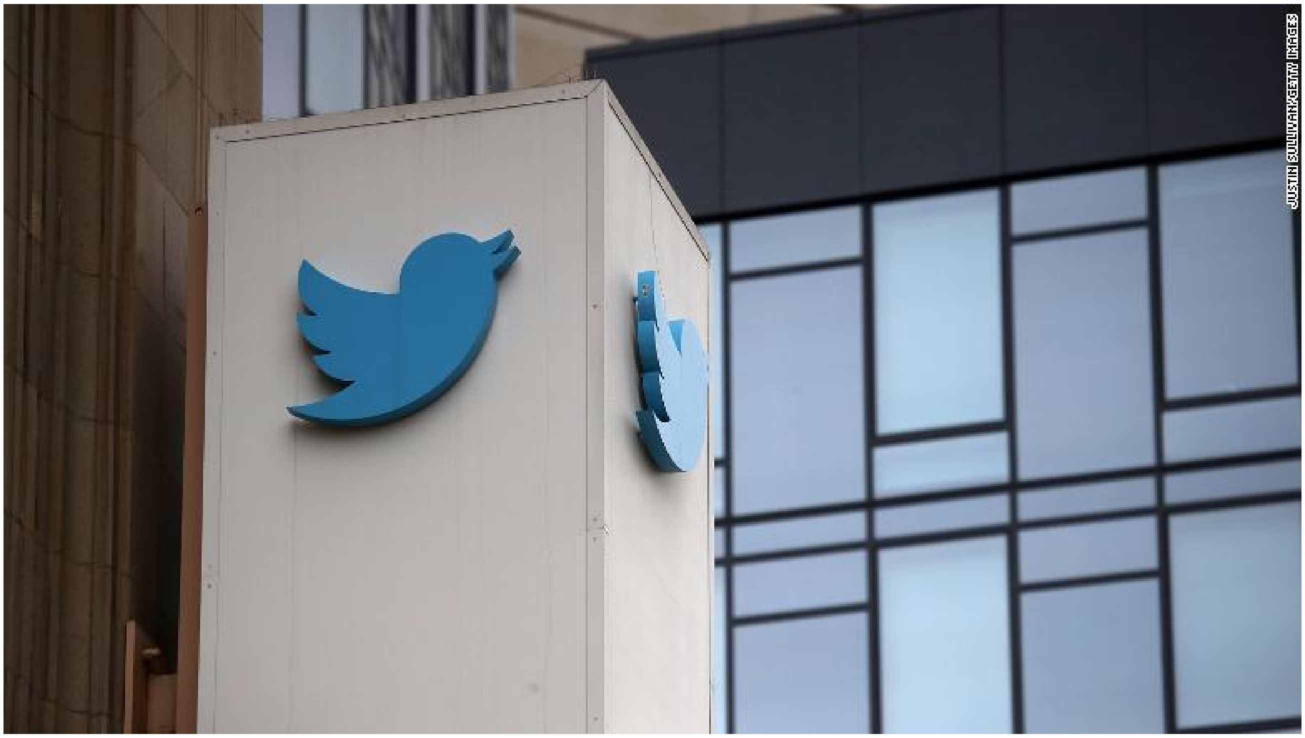
Bar Raisers - Large Group Justified

10/16/2020

DATA SCIENCE SKILLS THAT MATTER

Our group embarked on the quest to find an answer to the title question. Our approach consisted of trying to identify the terms that are commonly tagged along with the Twitter handle #Datascience. This provides the keywords that are most often associated with ‘data science’. The top 20 de-duped keywords will serve as a dictionary for the analysis. It is important to note that, these keywords may not be used in professional job listings. To validate the findings from Twitter, professional job listing sites in the US such as LinkedIn and Indeed will be used.

Twitter



Twitter headquarters

Twitter is a social network founded in 2006, it has over 325 million members and serves as a barometer of public opinion. Social media has played a fundamental role in the social activism of the 21st century. Nevertheless, people share their opinions and connect in a wide range of areas including careers. Is precisely this fact that motivated us to treat Twitter as a barometer for what hashtags people associate most commonly to Data Science.

Getting Twitter Data

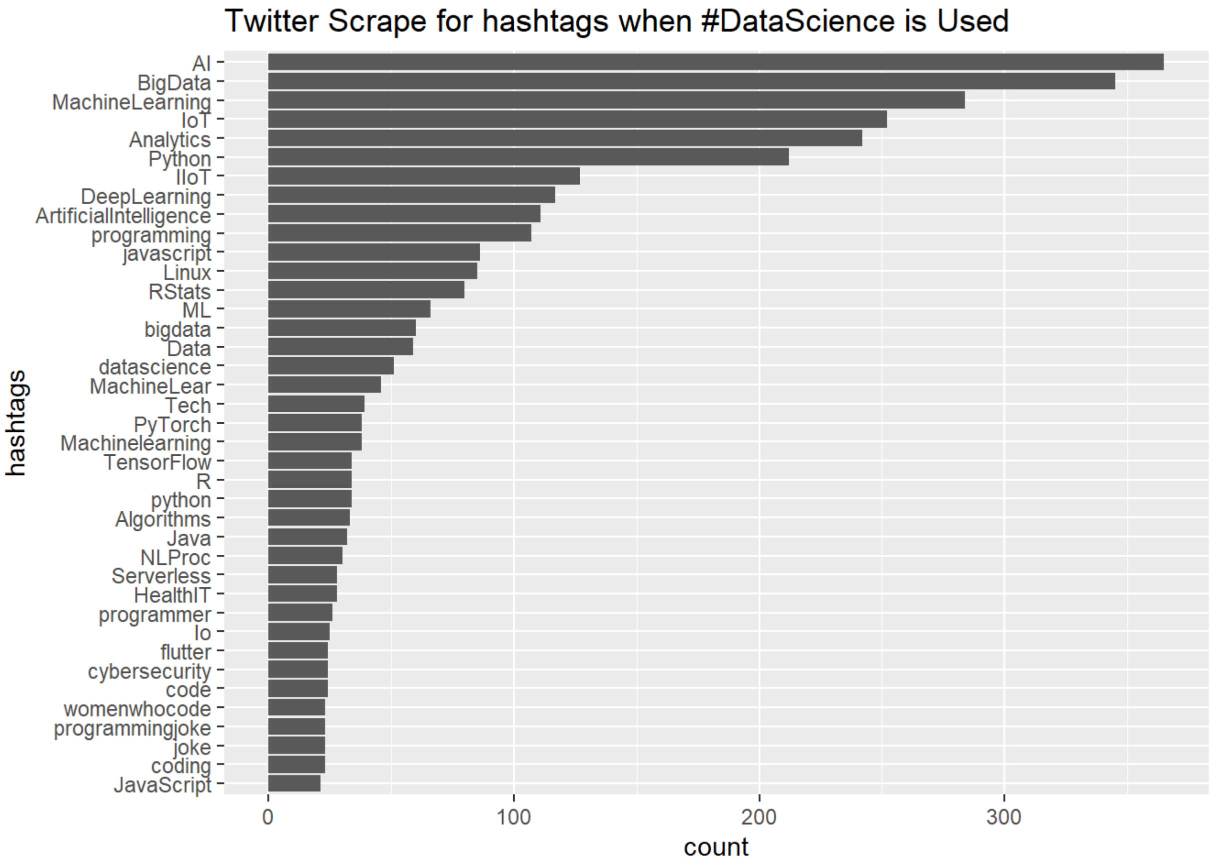
To have access to the Twitter data, we used the twitteR package. An API account and app had to be created to be able to access Twitter’s API. After connecting to twitter, we asked for the top 1000 tweets containing the hashtag #DataScience. We extracted the words from these tweets and got 5200 words. If we group and count the words that are repeated we get about 493 words.

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

word	word_count
AI	365
BigData	345
MachineLearning	284
IoT	252
Analytics	242
Python	212

Visualizing Common words

```
count_word_cut<-count_word%>%
  filter(word_count>20)

ggplot(count_word_cut, mapping=aes(x=reorder(word,word_count),y=word_count))+
  geom_bar(stat="identity")+
  coord_flip() +
  labs(title="Twitter Scrape for hashtags when #DataScience is Used",x="hashtags", y="count")
```



What we learned from Twitter

As we can see in this plot, the terms most people associate with Data Science are: - AI, Big Data, Machine Learning and Analytics

We see different languages associated with Data Science, of which *Python* appears to be number one.

We also see frameworks like Tensorflow and PyTorch.

LinkedIn

LinkedIn

Started in 2003, LinkedIn began as a social network for professionals and has evolved throughout the years into a Networking platform with career building capabilities, training and job search functionality. LinkedIn pivoted from a social network to a full fledged enterprise tailored to career searching, training and networking. Linkedin has topped at 315 million members and, according to recent statistics, the number of business professionals in the world is estimated between 350 and 600 million individuals. This means that over 50% of the business professionals in the planet are on LinkedIn!* See source (https://thelinkedinman.com/history-linkedin/#:~:text=LinkedIn%20started%20out%20in%20the,the%20New%20York%20stock%20exchange.)

Being able to query the job listings in LinkedIn would be an invaluable resource in answering the proposed question. Scraping data from LinkedIn used to be easier a few years ago but LinkedIn has implemented some security measures to ensure its members data is kept safe. At the same time, they have made scraping job postings harder. Nevertheless, scrape we shall.

Getting LinkedIn Data

To get the data from linkedin, we wrote a set of linkedin scraper functions using rvest. You can find these functions within the `functions` folder:

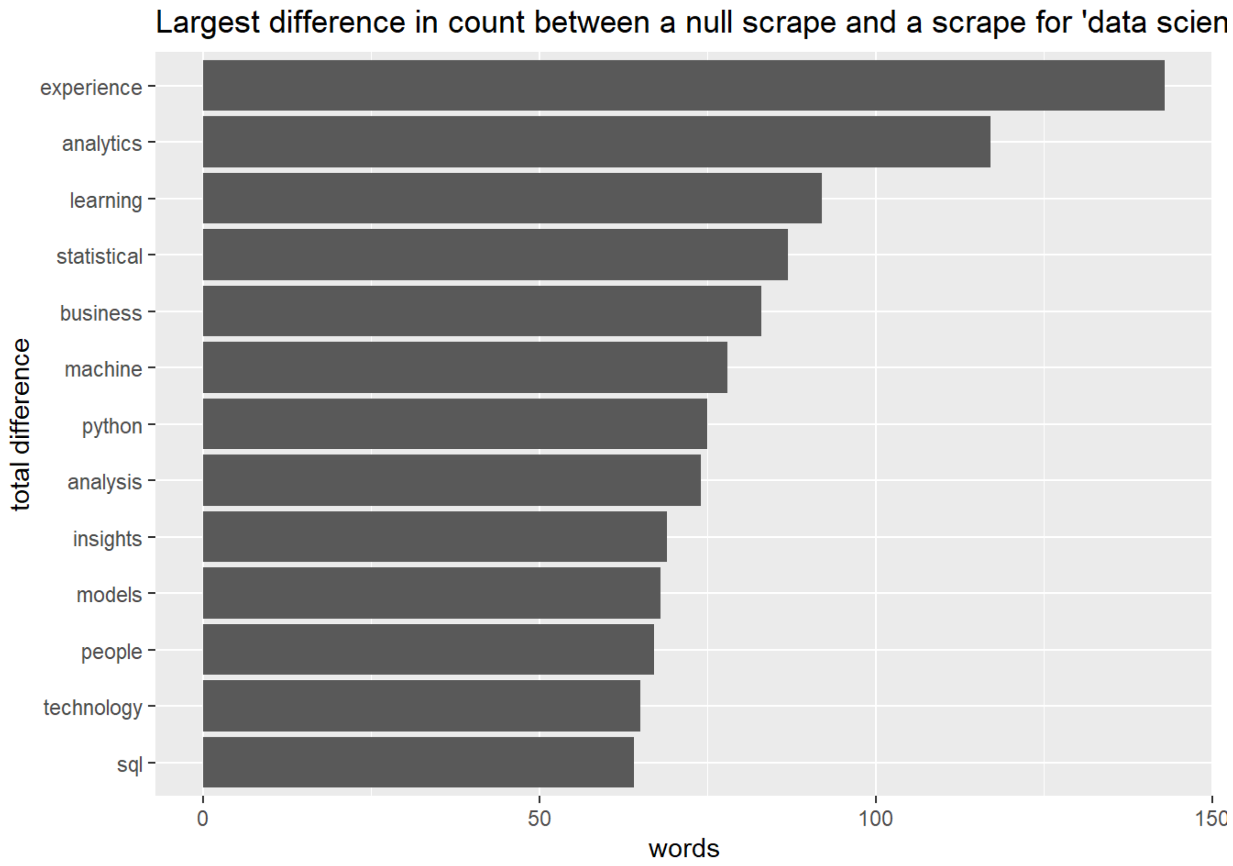
```
# Function:       linkedIn_scrape
# returns:        data frame
# link_base_url:  is the base url without the appended page number thing
# max:            is the maximum number of pages you want to scrape
# all_links:      returns all the links found
# all_links_unique: ensures the job links are unique
# all_jobs_scrape:iterates thru the links and scrapes the data

linkedIn_scrape<-function(link_base_url,max){
  link_base_url<-link_base_url
  max<-max
  job_link_list<-all_links(link_base_url=link_base_url,max=max)
  final_output_df<-all_jobs_scrape(job_link_list)
  return(final_output_df)
}

linkedIn_scrape_unique<-function(link_base_url,max){
  link_base_url<-link_base_url
  max<-max
  job_link_list<-all_links_unique_check(link_base_url=link_base_url,max=max)
  final_output_df<-all_jobs_scrape(job_link_list)
  return(final_output_df)
}
```

Null Scrape vs Data Science Scrape

For this analysis, we scraped the data to 250 job postings in Linkedin, by examining their description, we looked for the most common used words in the job posting. We also wanted to compare these results to what we will get if we did not search specifically for the keyword “Data Science”.

word	word_count
AI	365
BigData	345
MachineLearning	284
IoT	252
Analytics	242
Python	212

We then transformed this dataset, which at this point consisted of **10,132** observations. For the Null Scrape vs Data Science Scrape, we used the term count difference between the two datasets and prepared the most common ones to be examined by a plot.:



Looking at this graph, we see words that we did not see in our twitter search:

-**experience and people** appear to be pointing to Soft Skills as opposed to technical skills. -**Analytics, Machine Learning and Statisics** also appear to be valued more than specific frameworks or packages. -**Python** once again appears to be the top language, in this case, followed by SQL.

Geo Focused LinkedIn Analysis (Data Scientists positions in Brooklyn, NYC)

Job_Title	Company_Name	Company_Location	Num_Apps
Data Scientist	Koneksa Health	New York, NY	62 applicants
Marketing Data Scientist	Google	New York, NY	35 applicants
ML/AI Data Scientist	Lemonade	New York, NY	30 applicants
Quantitative Researcher/ Data Scientist	FinTech Startup	New York, NY	47 applicants
Data Scientist	JPMorgan Chase & Co.	Jersey City, NJ	40 applicants
Point72 Data Scientist	Point72	New York, NY	56 applicants

Tidy the data

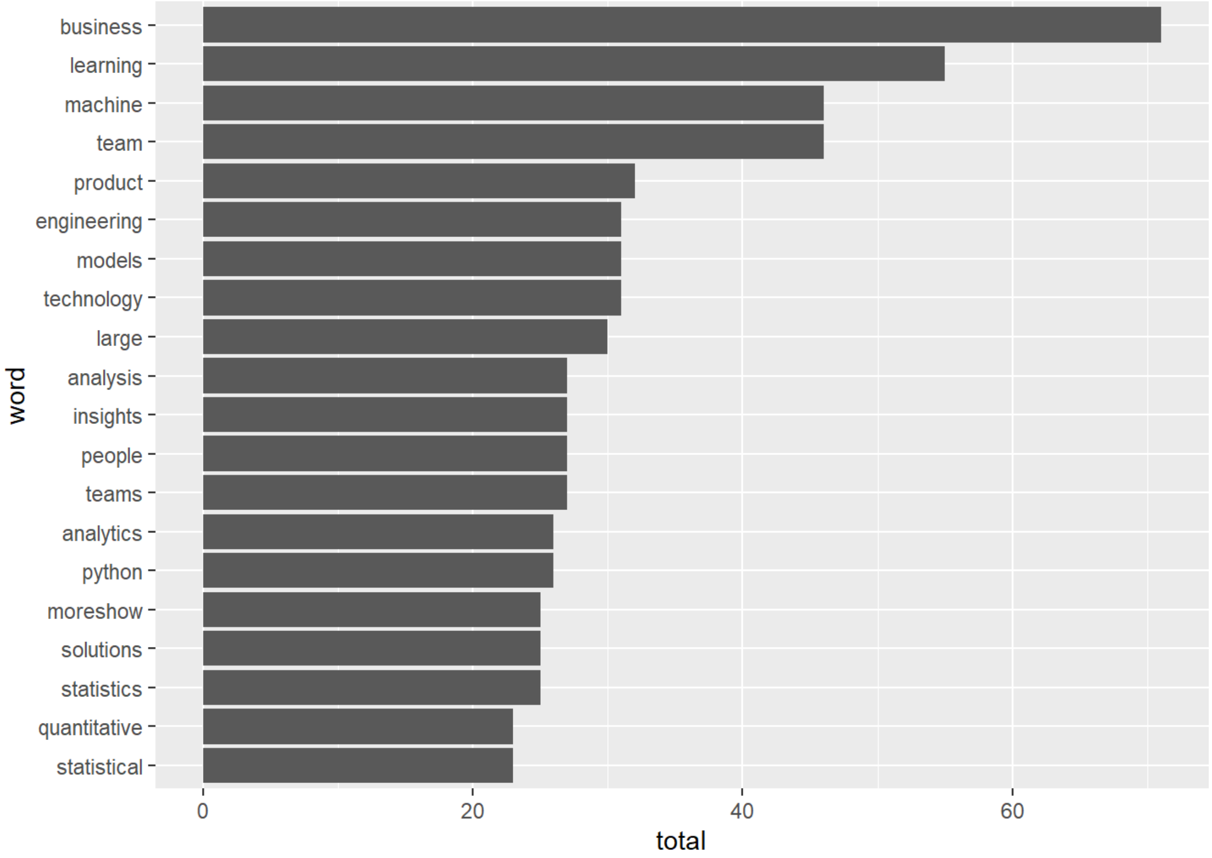
The job descriptions need to be tidied so we can analyze what companies are looking for in a data scientist. Tidying will include extracting and counting individual words, removing stop words such as “the”, “a”, and “and”.

Further tidying is performed by analizing combinations of two words “bigrams”, separate the bigrams, remove stopwords, perform a count and unite both words into a single column.

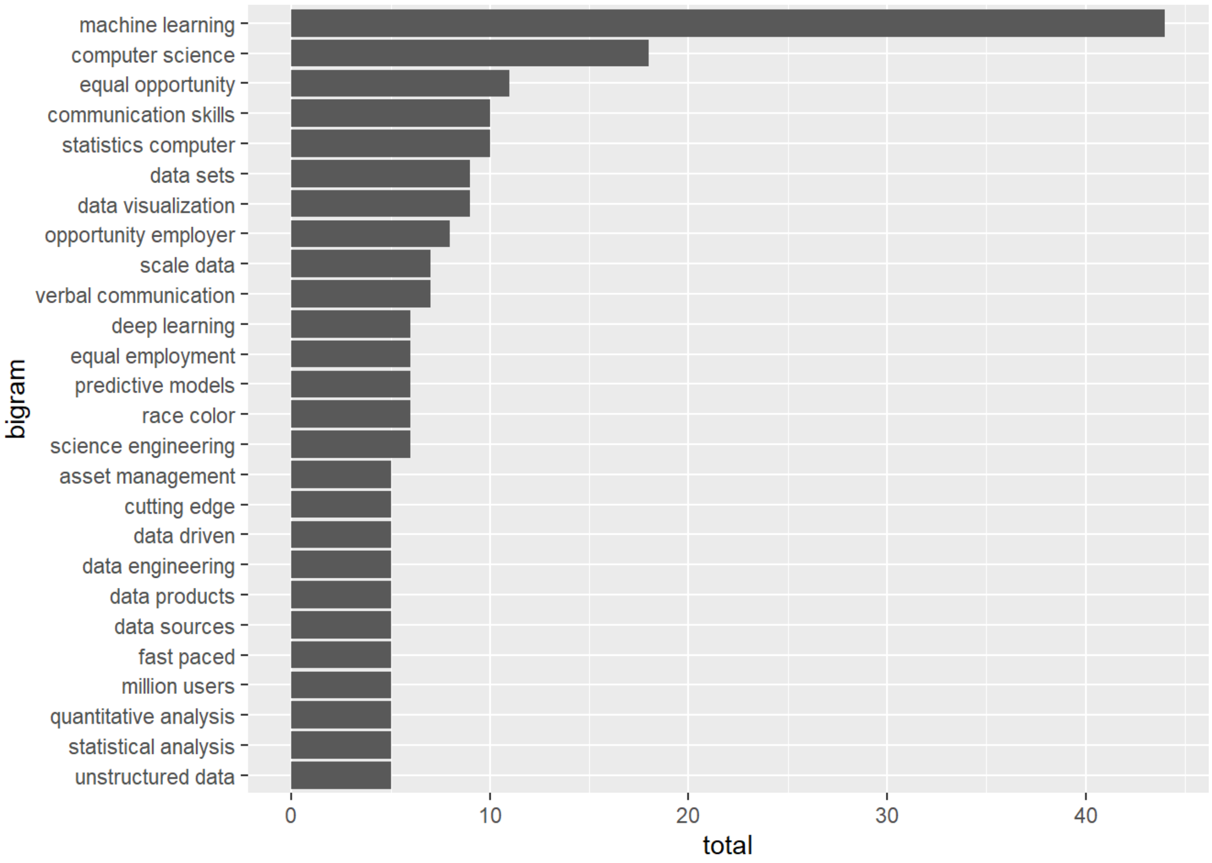
Job_Title	bigram	n
Data Scientist	09 28th	1
Data Scientist	1 3	1
Data Scientist	1 4	1
Data Scientist	10 business	1
Data Scientist	11 billion	1
Data Scientist	16 trillion	1

Analysis

The bar plots below look at the frequency of words in all the job descriptions. The words data, science, and scientist were excluded from this analysis since those are the words used in the job descriptions and it will skew that data. Other words such as years, and work were also removed. The first chart looks at the top 20 single words. It is easy to point out which words are frequently used, but it does not really give us much information besides frequency.

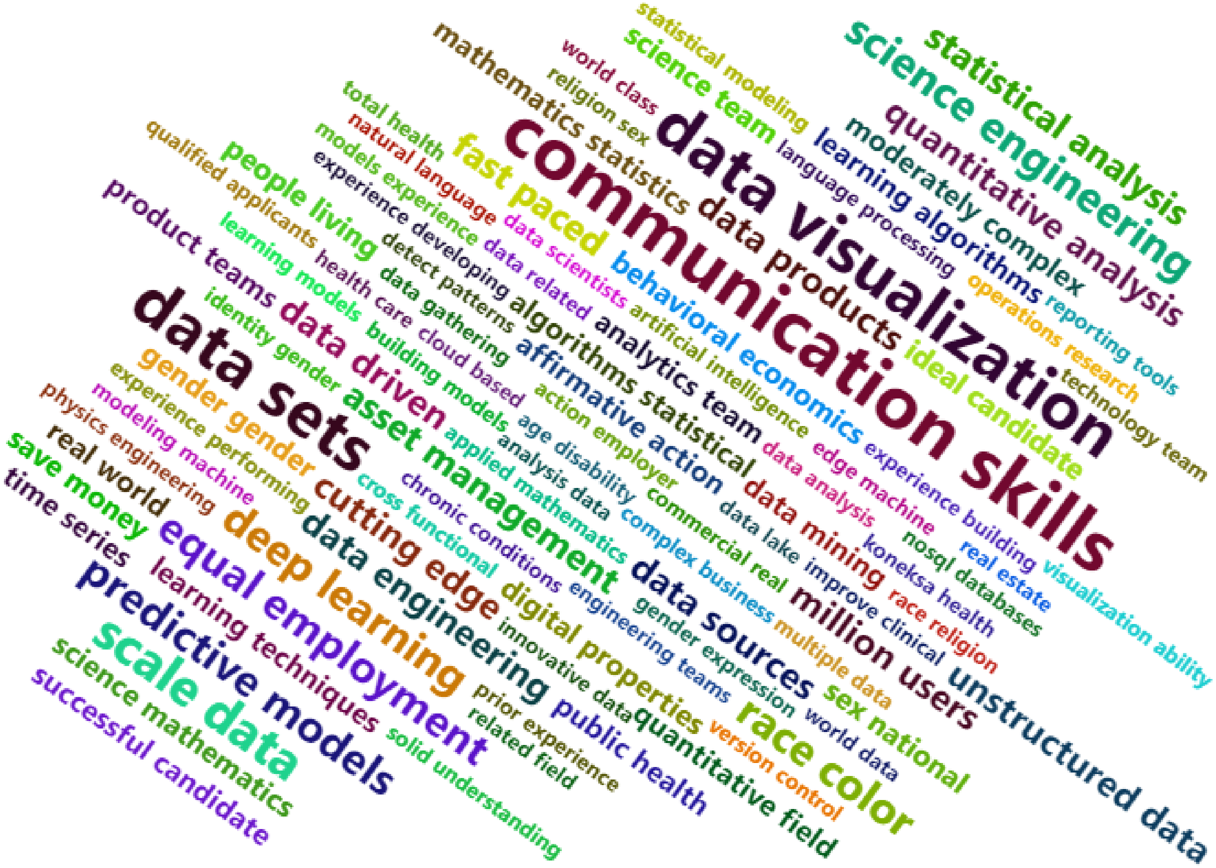


The bigram gives us more information on what employers are looking for and it looks like most frequently used bigrams are hard skills. Machine learning is the top skill out of all skills and happens to be a hard skill. The top soft skill in communication.



Visualization

We used a wordcloud to create a visualization of the top 50 words used in LinkedIn job descriptions for Data Scientist jobs. Computer Science is important to employers along with communication skills, data visualization, and predictive models.



Amazon

So far, we have analyzed a sample data set from LinkedIn and Twitter to understand the trends in ‘Data Science’. Now, let’s take a large dataset containing all the jobs related to ‘Data Science’ to validate our hypothesis. For this purpose, we need to select a company which is one of the largest hiring company for data science related roles and also is a reputable company that has positive reviews in the job market.

One of the largest job portals company, glassdoor.com (https://www.glassdoor.com/blog/companies-hiring-data-scientists/) shows **‘Amazon’** as the best candidate for this purpose.

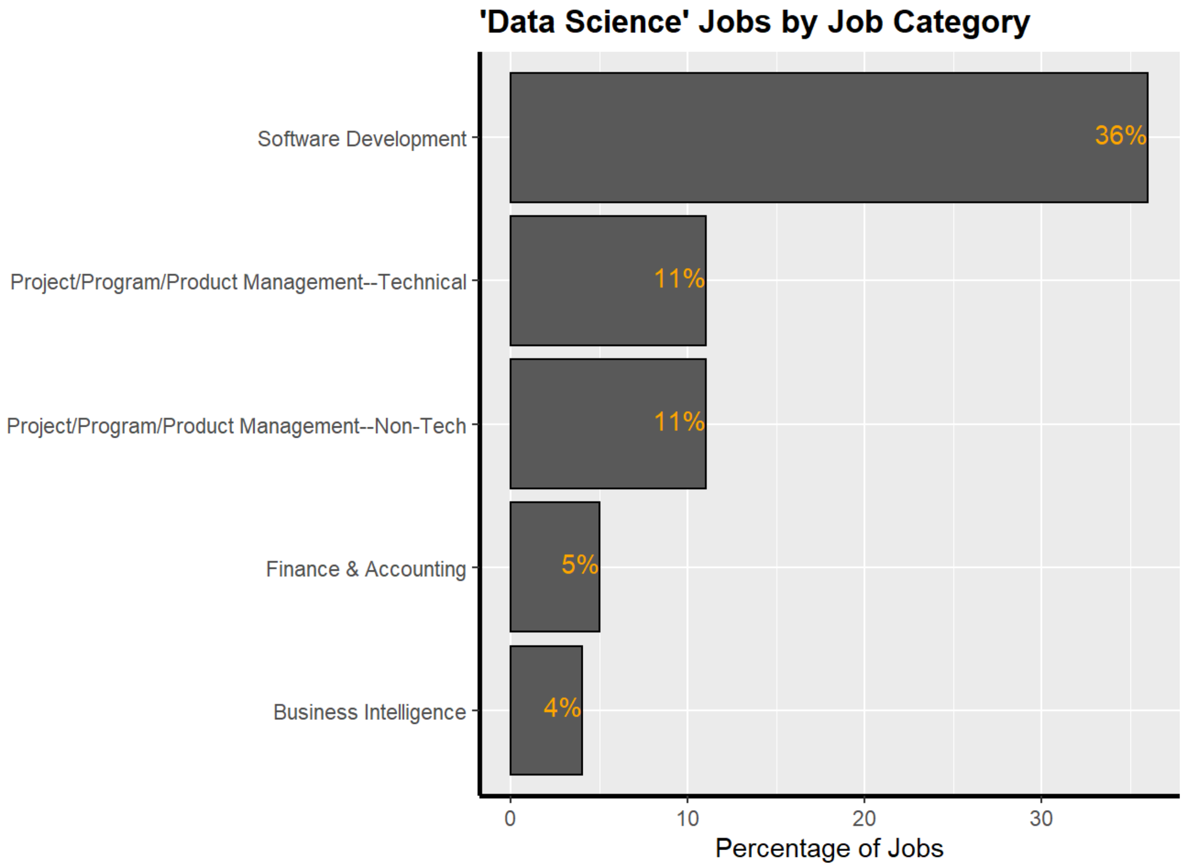
The search on Amazon.jobs (https://www.amazon.jobs/en/search?base_query=Data+Science&loc_query=United+States&type=area&latitude=38.89037&longitude=-77.03196&country=USA) for keyword “Data Science” in the location “United States” revealed 6630 results. This data is scraped from the website using python query and using a proxy website ‘Crawlera’.

Tidying data for Analysis

Select only the required columns to perform the analysis. Both ‘Basic Qualifications’ and ‘Preferred Qualifications’ appear to be key columns(apart from job category), let’s select them both and subset the data to the year **2020** alone.

Data reveals that **‘Software Development’** department offers the most data science jobs at Amazon! Let’s look into the table a bit. In total, the top three job categories such as **Software Development**, **Product Management (technical+non-technical)** and **Finance & Accounting** contribute to **63%** of the data science jobs.

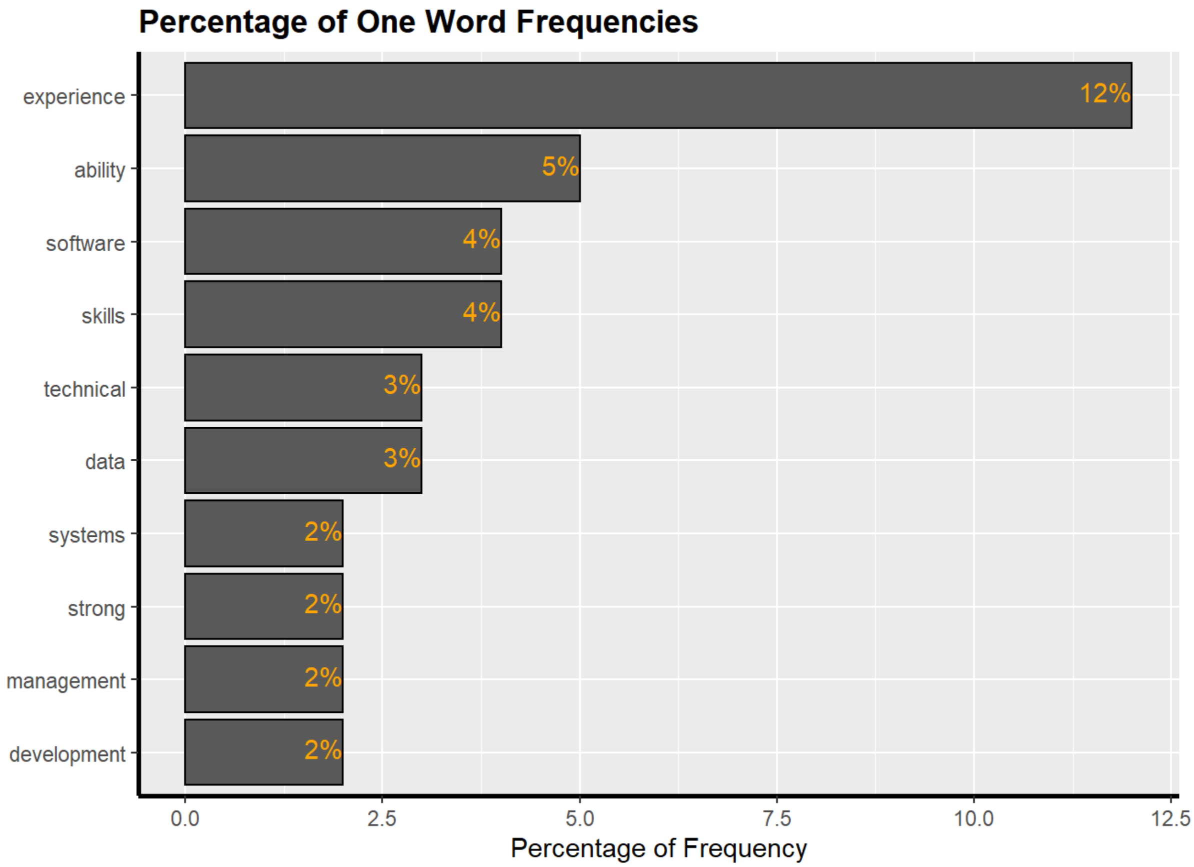
Job.category	n	jobs_pct
Software Development	2177	36
Project/Program/Product Management–Technical	669	11
Project/Program/Product Management–Non-Tech	649	11
Finance & Accounting	274	5
Business Intelligence	233	4
Systems, Quality, & Security Engineering	224	4
Solutions Architect	208	3
Operations, IT, & Support Engineering	147	2
Sales, Advertising, & Account Management	141	2
Design	127	2



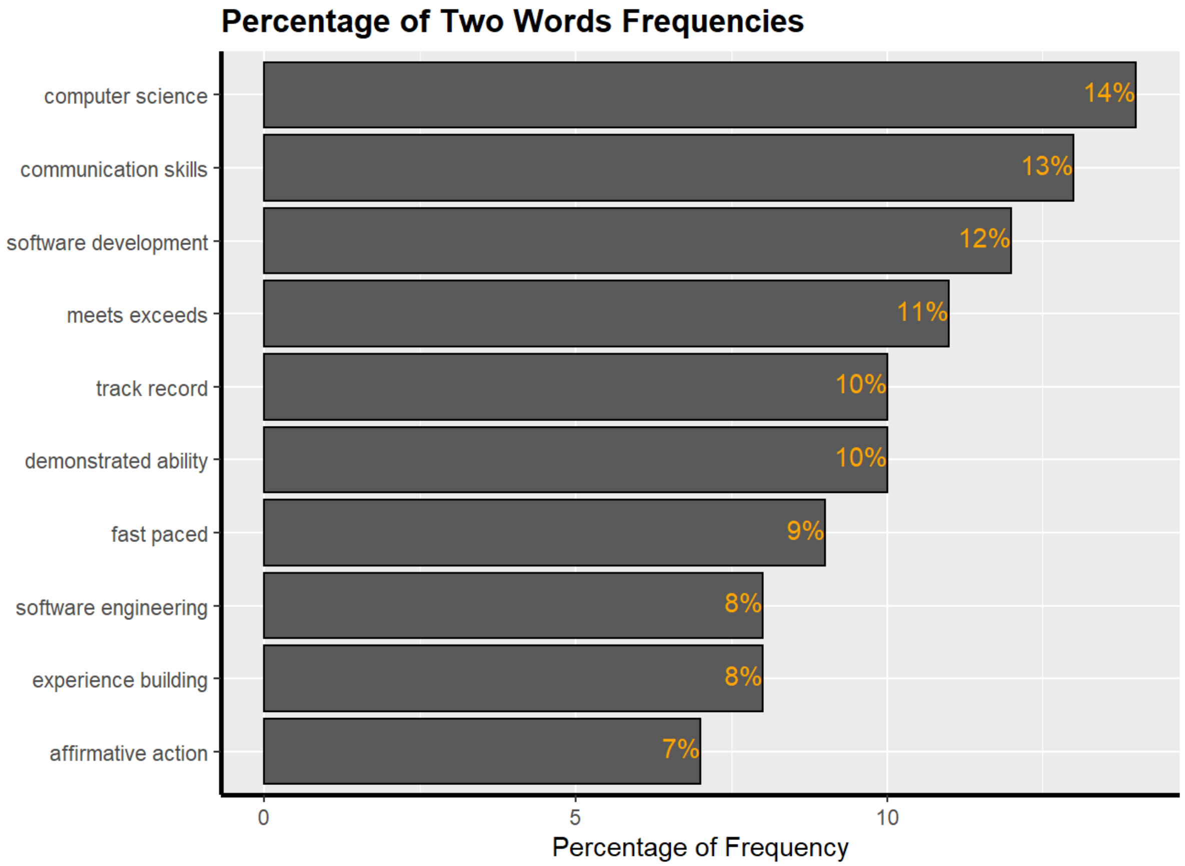
On closer look, ‘Perferred Qualifications’ seen to have more information about the job than ‘Basic Qualifications’. And next, removing legal and PR content from the column to perform text analysis

Exploratory data analysis

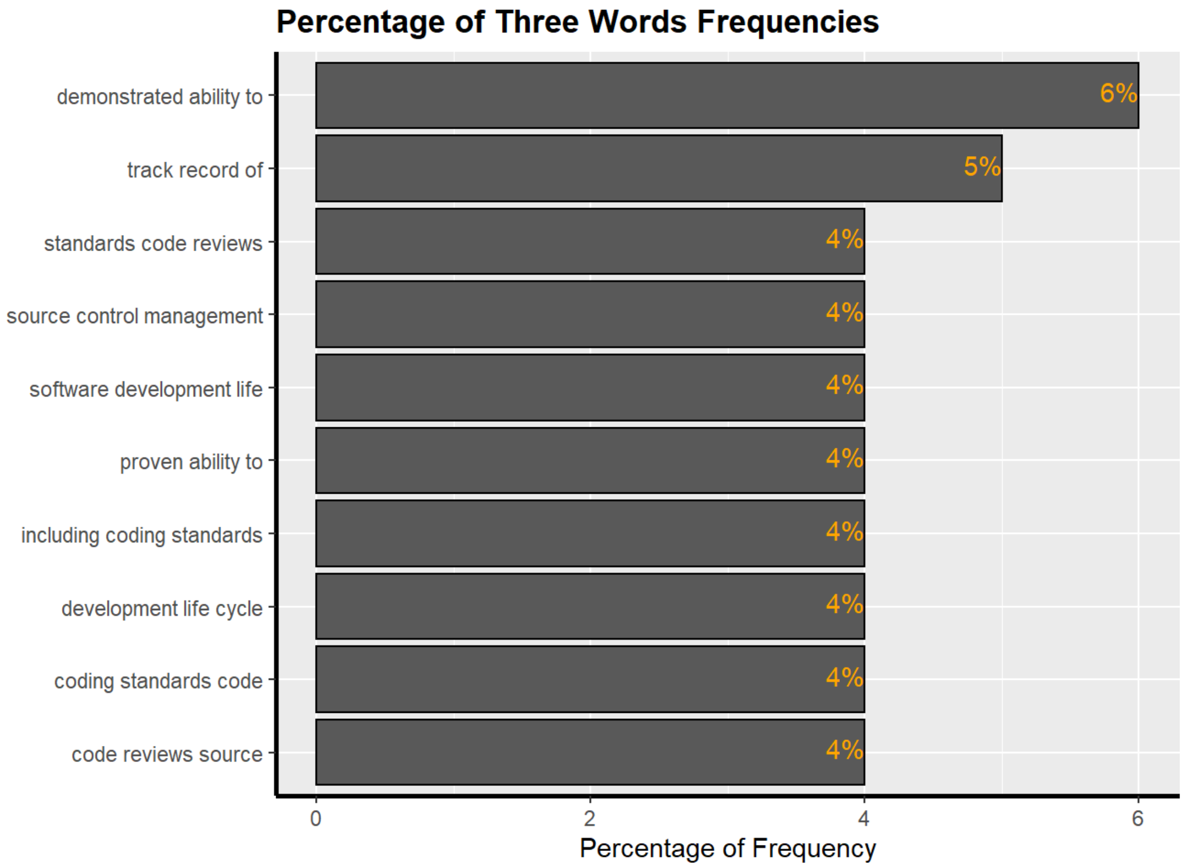
One Word Frequency count



Two Words Frequency count

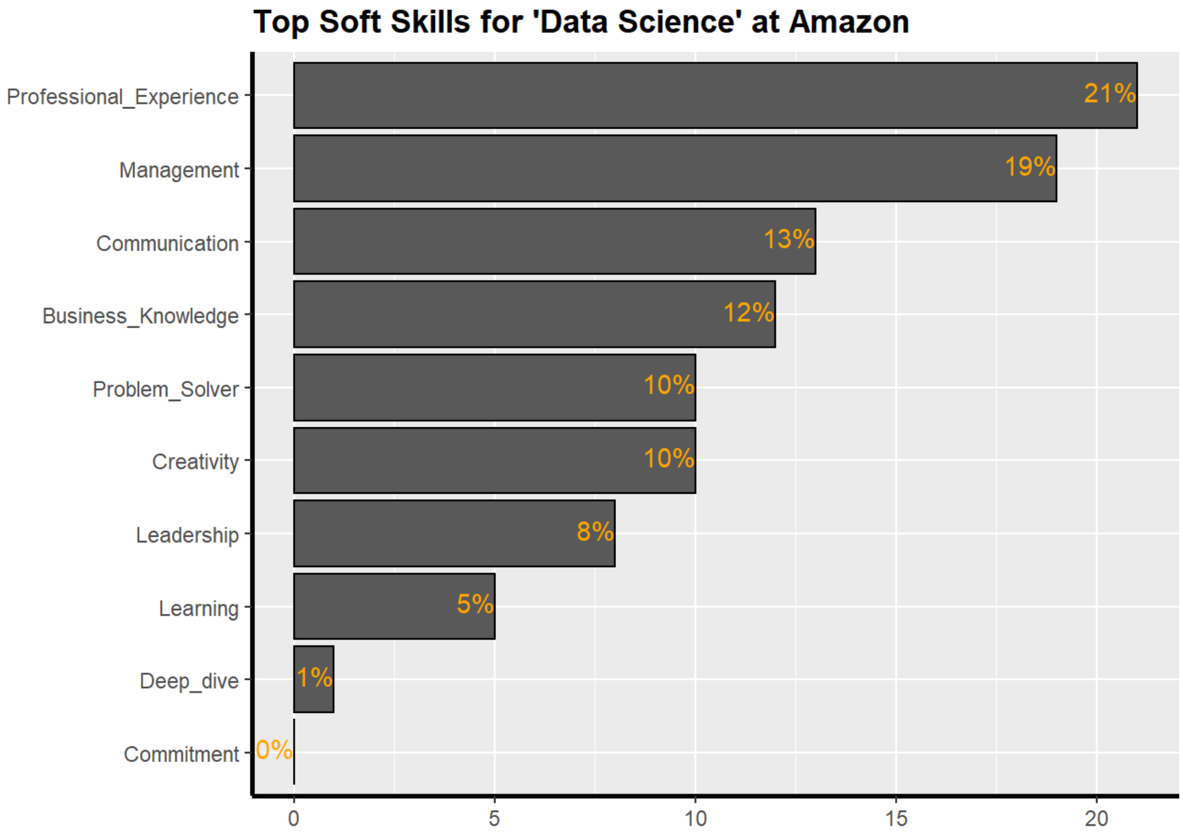


Three Words Frequency count

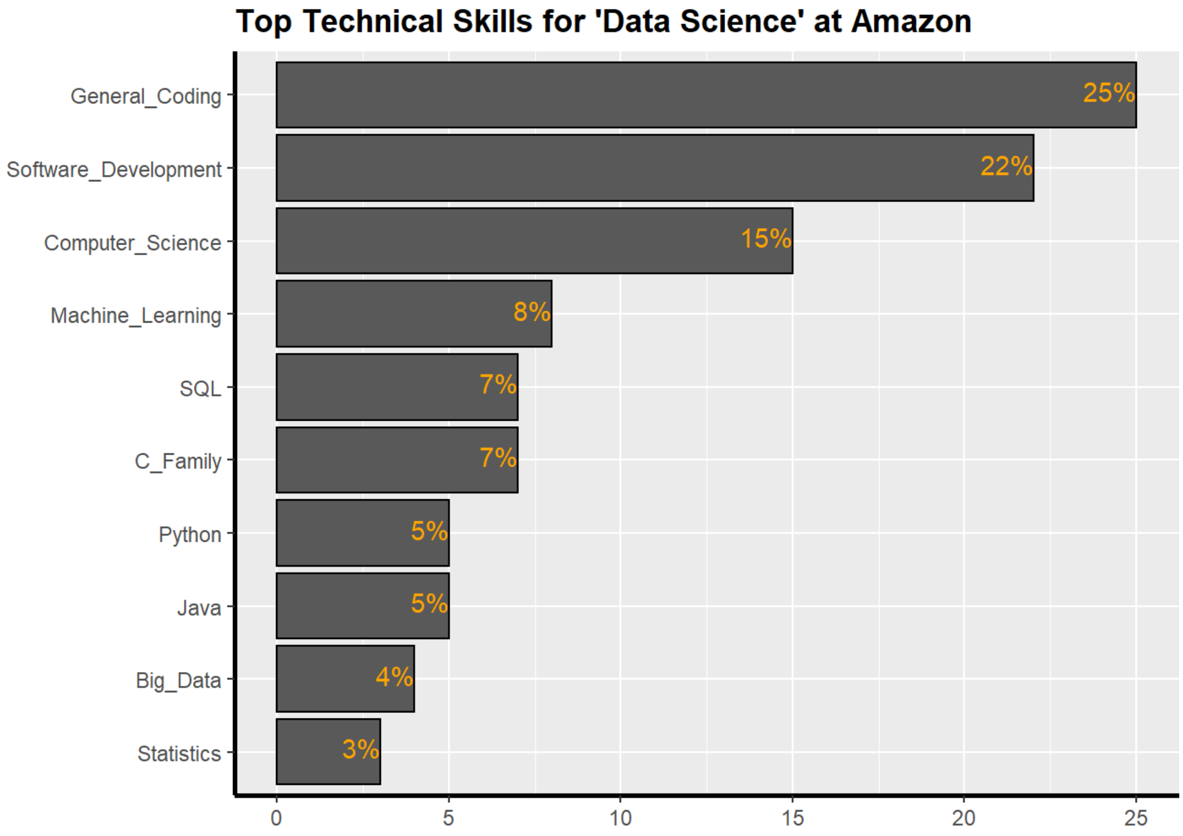


Creating dictionaries based on the observed “One Word”, “Two Words” and “Three Words” frequency counts. These trends are used to create two different sets of skills - **Technical skills** and **Soft skills**. Thus, we can conclude that statistics is both science and art!

The most important **Soft skills** are:

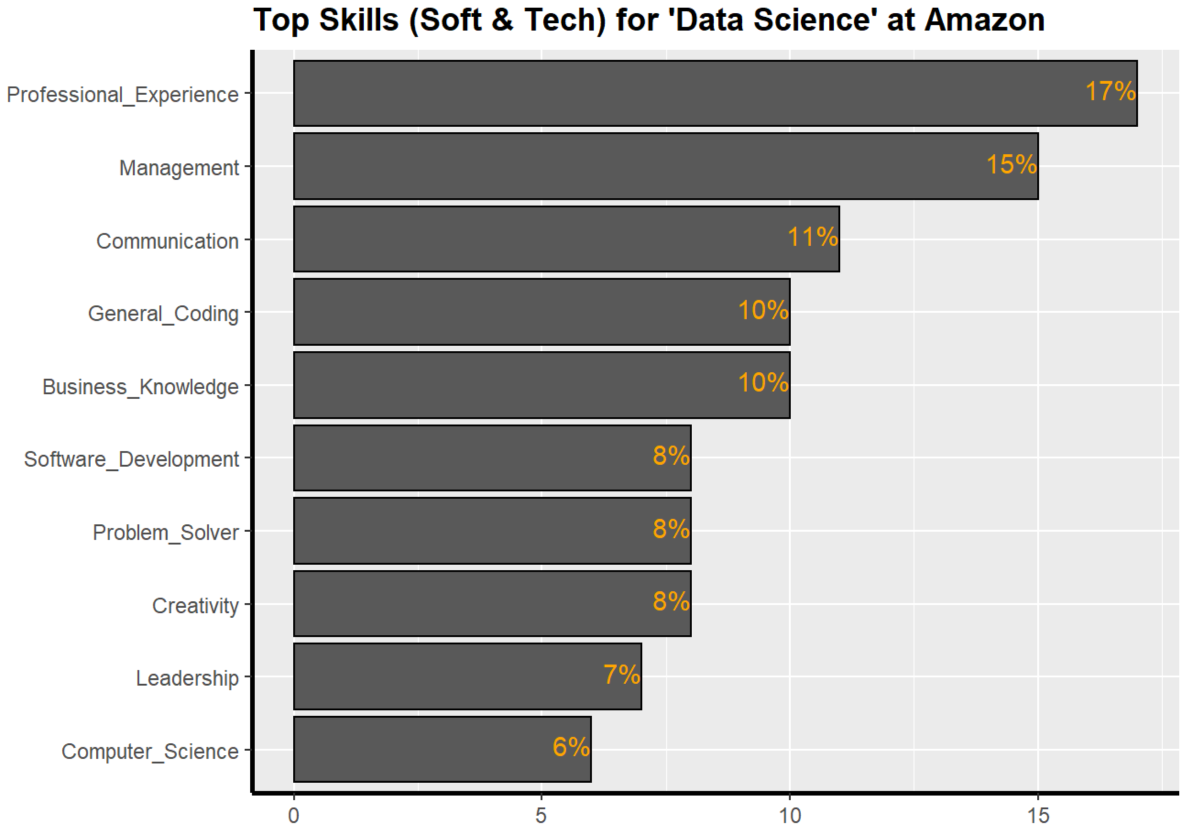


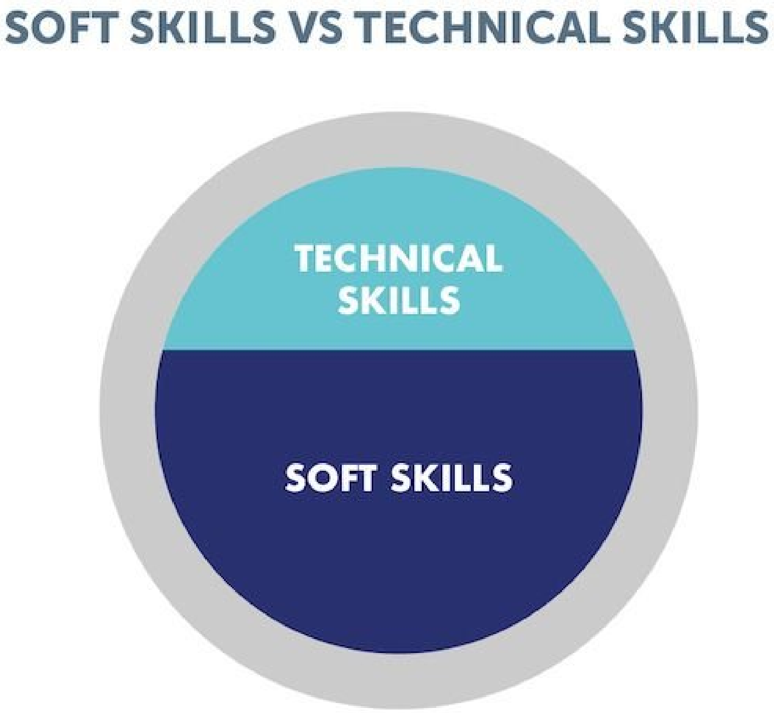
The most important **Technical skills** are:



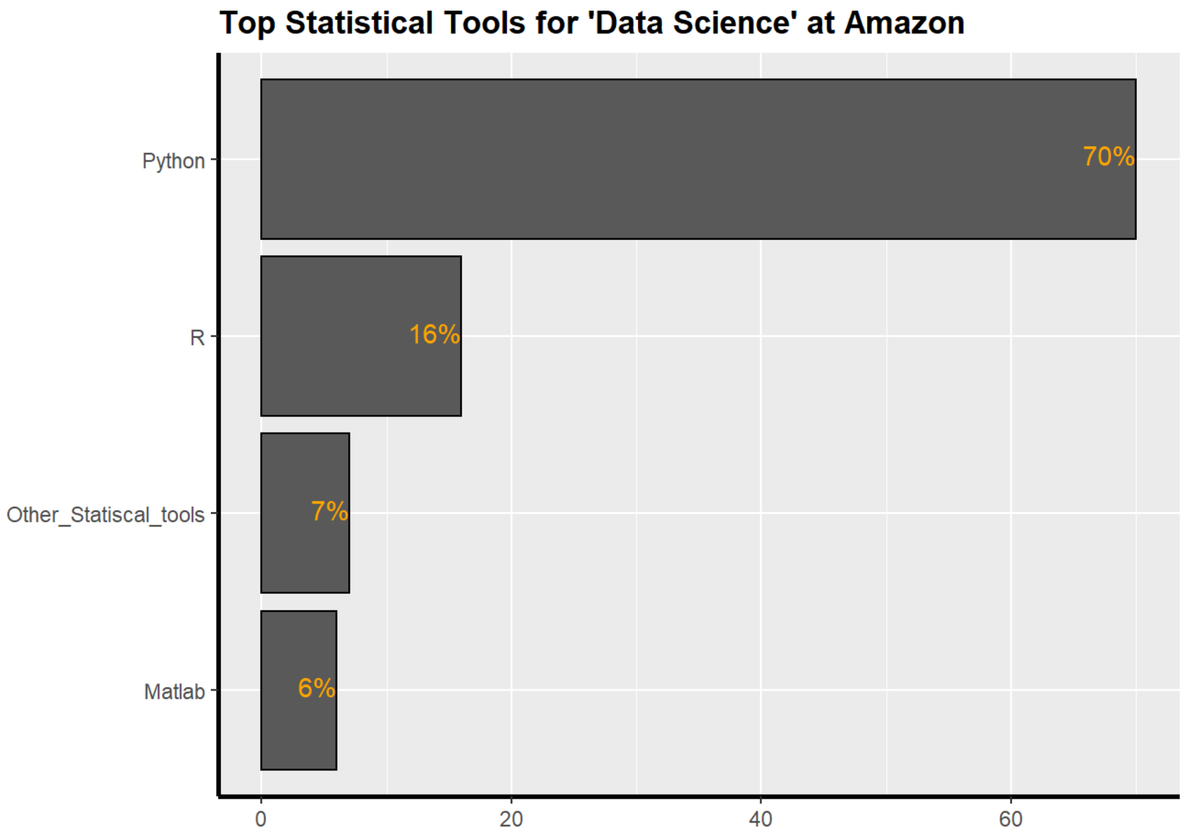
Conclusion

Now, let’s combine both soft skills and technical skills and find the top Data Science skills at **Amazon**. We find that **8 out of top 10** skills required at Amazon are **Soft Skills**. So, when you prepare for an interview at Amazon next time, be sure to focus on highlighting your **“Soft Skills”**





Bonus: Let’s find the most popular statistical tool among **Python**, **R**, **Matlab** and others at **Amazon**



Supporting Scripts

As part of this project we wrote several supporting scripts and functions that helped us organize, scrape, and make the overall project more efficient. Here we list some of them but you can find them all within the appropriate folders:

- *functions*. Holds common functionality
- *scraping*. Holds the scraping functions

AWS Connection Script

```r
library(DBI)
library(tidyverse)

getDbConnection <- function() {
  # This function returns a DB Connection to AWS
  password_file<-"C:\\password-files-for-r\\AWS_login.csv"

  # read in login credentials
  df_login<-read.csv(password_file)

  # Uncomment this next line to see the values
  # cat("Host: ", vardb_host, " Schema=", vardb_schema, " username=", vardb_user, " password=",
 vardb_password)
  mydb = dbConnect(RMySQL::MySQL(),
                   user=df_login$login_name,
                   password=df_login$login_password,
                   port=3306,
                   dbname=df_login$login_schema,
                   host=df_login$login_host)

  return(mydb)
}
```

LinkedIn Scraper

```r
# function: linkedIn_scrape
# returns data frame
# link_base_url: is the base url without the appened page number thing
# max: is the maximum number of pages you want to scrape
# all_links: gets all job links
# all_jobs_scrape: iterates thru all job links and scrapes jobs

linkedIn_scrape<-function(link_base_url,max){
  link_base_url<-link_base_url
  max<-max
  job_link_list<-all_links(link_base_url=link_base_url,max=max)
  final_output_df<-all_jobs_scrape(job_link_list)
  return(final_output_df)
}

linkedIn_scrape_unique<-function(link_base_url,max){
  link_base_url<-link_base_url
  max<-max
  job_link_list<-all_links_unique_check(link_base_url=link_base_url,max=max)
  final_output_df<-all_jobs_scrape(job_link_list)
  return(final_output_df)
}
```

Credits

- Bharani Nittala
- Richard Sughrue
- LeTicia Cancel
- George Cruz Deschamps

- Jack Wright

…