A Closer Look into the Skills in Demand for Data Scientists, Data Engineers, and Data Analysts

Yini Rong

Table of Contents

[TL;DR 2](#_Toc132884029)

[Introduction 2](#_Toc132884030)

[Problem 2](#_Toc132884031)

[Goal 3](#_Toc132884032)

[Data Collection 3](#_Toc132884033)

[1) Web scraping 3](#_Toc132884034)

[2) Data limitations 4](#_Toc132884035)

[Analysis 4](#_Toc132884036)

[1) Skill keyword identification 4](#_Toc132884037)

[2) Skill keyword frequency 5](#_Toc132884038)

[a. Fuzzy match algorithm 5](#_Toc132884039)

[b. Word clouds 7](#_Toc132884040)

[3) Skill importance 9](#_Toc132884041)

[a. TF-IDF calculation 9](#_Toc132884042)

[b. Unique skills 9](#_Toc132884043)

[4) Context on skills 10](#_Toc132884044)

[a. Gensim 10](#_Toc132884045)

[b. Variations in common skill applications 10](#_Toc132884046)

# TL;DR

[[3 sentences or bullet points on the most interesting points of my project]]

# Introduction

A myriad of job fields with an emphasis on working with data has been growing significantly across various industries in the past decade. Careers, such as data scientist, data engineer, and data analyst, are becoming increasingly popular. As data industry jobs continue to evolve, it is essential to understand the skills that are most in demand for these roles so that job seekers can best prepare for their desired career trajectories. My latest project aims to address this by analyzing thousands of job postings for data scientist, data engineer, and data analyst positions using text analytics.

# Problem

Despite the decrease in data scientist hiring due to widespread hiring freeze in the high-tech industry, the job market for data engineers and data analysts is expected to sustain growth over the next few years.[[1]](#footnote-1) Demand for these positions has continued to rise, fueled by the increasing reliance on data-driven decision-making in various industries.

By shedding light on the skills that are most in demand for data industry jobs, this project offers valuable insights for job seekers who are either seasoned professionals or just starting out. My project can help better understand the skills that are essential for success in these data-centric fields.

# Goal

While previous studies have attempted to identify the most important skills for data industry jobs[[2]](#footnote-2), this project offers a fresh perspective and updated insights based on the most recent job postings available. By connecting several fundamental techniques in text analytics, my project aims to present a comprehensive analysis of the frequency and context of the most important skills across the three roles – Data Scientist, Data Engineer, and Data Analyst.

# Data Collection

## Web scraping

I collected 2,908 jobs posted on LinkedIn in March 2023 through web scraping. I focused on scraping full-time entry-level job posts in the United States with the following titles: “Data Scientist,” “Data Engineer,” “Analyst.” While the Data Scientist and Data Engineer job posts have the strings, “Data Scientist” and “Data Engineer,” respectively in the titles, I decided to include analyst job posts that have titles other than “Data Analyst,” such as “Business Analyst” or “Business Intelligence Analyst.” This over-inclusion is to recognize the broader nature of an analytics role in hope to capture a more accurate set of skills required for being an analyst who works with data. Despite the over-inclusion, job posts with “Data Analyst” as the title consist of [[80?]] of all analyst job posts.

For each job post, I collected data from the following eight fields: job post ID, date posted, company, job title, location, industry, function, and job description.

[[Insert table showing for each job title, # of job posts, date posted range, sample job titles]]

## Data limitations

The data I collected is subject to biases. The first is sample bias. Web scraping was a time-intensive process, so I collected data from only one job search engine. Since LinkedIn is perceived as the most popular network for recruiting for professional roles[[3]](#footnote-3), it is a good starting point for obtaining a decent set of posts in the fields of data science, data engineering, and data analytics. However, dozens of alternative job search engines, including Indeed, ZipRecruiter, and Glassdoor, might have job posts that were not posted on LinkedIn. As a result, my dataset might not be as comprehensive as it could be in capturing all job posts of interest.

Secondly, there is recency bias in my dataset. LinkedIn only has job posts going back one month; therefore, insights derived from this dataset must be contextualized in the condition of the job market and the hiring needs of a limited set of companies in March 2023.

# Analysis

[[Short summary on the analyses done]]

## Skill keyword identification

To understand the purpose of this step, one needs to understand the challenge in the data available. If one simply creates a word cloud using all Data Scientist job posts, one will not get much insightful information, as the word cloud will show “experience” as the biggest keyword and many that are too broad or vague, such as “support” and “ability.” It is not hard to imagine that recruiters would want to attract candidates who have relevant experience and the ability to perform on the job. Data from the job descriptions is full of noise. Instead of removing trivial keywords from the job descriptions, I decided to define a list of key skills needed for these positions.

Using a random sample of the data, I complied a master list of 142 skills that show up in the job descriptions. Skills were extracted only if they could be measurable in a meaningful way. That is, these skills refer to the technical knowledge and capabilities that companies look for in Data Scientists, Data Engineers, and Data Analysts. Some examples of these keywords are “machine learning,” “python,” and “database management.” Moreover, these skills could be objectively assessed through tests, or even by looking at specific work experiences or accomplishments.

Unsurprisingly, many soft skills, such as “communication,” “problem solving,” and “project management,” were listed in the descriptions. While these are certainly important skills, it is difficult to measure them in a consistent way across a range of industries. These soft skills are subject to different standards due to the differences across companies. As a result, I decided to focus on the indicators for technical capabilities and created a more objective list of keywords that could be used to assess what companies were looking for in their candidates.

## Skill keyword frequency

With the master list of skill keywords, I used a matching algorithm to find whether a keyword shows up in a job description. Then, for each of the three positions, I summed up the frequency of each skill keyword and created a word cloud visualizing skills that are most mentioned in job posts.

### Fuzzy match algorithm

The token\_set\_ratio function in the FuzzyWuzzy package is a useful tool for fuzzy string matching in Python. This function compares two strings and returns a similarity score based on the edit distance between them. The score ranges from 0 to 100, with higher scores indicating greater similarity.

Given two strings, the token\_set\_ratio function first tokenizes each of them by eliminating non-alpha, non-numeric characters, such as punctuations, and converting all tokens into lowercase. It then sorts the tokens. It also removes duplicates in each string, which means that it uses a set-based approach, considering only unique tokens. In many job descriptions, some words are repeated often, such as “learning,” “skill,” etc. The removal of duplicates helps to avoid overemphasizing tokens that have multiple instances in a string.

The tokens are split into two groups: the intersection set and the remainder set. Similarity scores are calculated for three comparisons, and the maximum is taken as the result.[[4]](#footnote-4) For example, consider the following two strings:

s1 = “machine learning techniques”

s2 = “demonstrated experience in data science and machine learning theory”

These two strings are then split into intersection and remainders.

intersection\_set = “learning machine”

remainder\_set\_1 = “learning machine techniques”

remainder\_set\_2 = “and data demonstrated experience in science theory”

Then, the edit distance is calculated using the ratio function. The token\_set\_ratio function takes the maximum score among the three comparisons.

fuzz.ratio(intersection\_set, remainder\_set\_1) = 74

fuzz.ratio(intersection\_set, remainder\_set\_2) = 33

fuzz.ratio(remainder\_set\_1, remainder\_set\_2) = 39

fuzz.token\_set\_ratio(s1, s2) = 74

As we can see, if a particular skill is simply a word, e.g., “python,” or completely contained in the job description, e.g., “data pipeline,” the similarity score is close to 100, resulting in a match.

Though flexible in comparing strings that contain different word orders and additional words, the token\_set\_ratio function is aggressive in predicting a match. In other words, if any similar part of a skill keyword shows up in the job description, the function is likely to assign a high score to that keyword. For instance, “data management” can be assigned with a high score close to 100 if both tokens, “data” and “management,” appear in the description even when the concept or techniques of data management are not actually mentioned. Thus, minimizing false positives, or maximizing the precision score, is of interest here.

Due to the aggressive nature of the match function in predicting, I decided to choose a score of 70 as the cutoff threshold – considering a match if the similarity score is greater than 70 and not a match if below. [[I also performed an assessment on the precision score by varying the threshold from 40 to 80 using a set of labelled data. This assessment confirmed that a score of xx as the cutoff is optimal in minimizing false positives.]]

### Word clouds

While each role has different responsibilities and domain expertise, they often use similar tools and technologies to accomplish their goals. In the word clouds, we can see that Python and SQL are programming languages often mentioned across the three roles. Having a computer science background appears to be desirable for these roles as well.

On the other hand, we can see major differences in key skills across the three positions. For professionals who have experience in the data industry, these differences should not be surprising. Figure 1 emphasizes “statistical method,” “machine learning,” “deep learning” as data scientists are responsible for using these techniques to develop predictive models. Figure 2 highlights tools like “AWS,” “Spark,” “Azure,” used by data engineers to build and manage data pipelines and ETL processes. Lastly, data analysts are responsible for cleaning and analyzing data, and delivering insights to stakeholders. Figure 3 puts the emphasis on data visualization tools like “Excel,” “Tableau,” and “Power BI” to communicate their analyses.

Figure 1. Data Scientist Key Skills

Text

Description automatically generated

Figure 2. Data Engineer Key Skills

Text

Description automatically generated

Figure 3. Data Analyst Key Skills

Text

Description automatically generated

## Skill importance

Using TF-IDF can explicitly demarcate the set of skills unique to each of the three positions. Doing so could help job seekers in identifying the areas they need to improve on to gain a competitive edge in the candidate pool and to achieve the next step of their desired careers.

### TF-IDF calculation

[[Describe how I adapted the TF-IDF calculation to compute the importance of each skill.]]

### Unique skills

[[Describe the calculation of threshold to determine which skills are unique to which position.]]

There are 12 skills unique to Data Scientist position. There are close to 20 skills unique to Data Engineer position. However, Data Analyst does not have many unique skills from the former two positions. This could indicate that the skillsets needed for Data Analyst are required for Data Scientist and Data Engineer as well.

[[Present the table below in a more digestible format]]



## Context on skills

Using Word2Vec to see how the clusters of the top 20 similar words for common skills ***differ*** across DS, DE, and DA. This is to understand the relationships between skills and how these skills being applied differently for the three positions.

### Gensim

[[Describe broadly how I used genism (hyperparameters, normalized vectors) and the most\_similar function to get top similar words for common skills]]

### Variations in common skill applications

[[Pick a few common skills, e.g., Python, based on common importance calculation from TF-IDF and show examples of the different top similar words for each of the three positions]]

[[Show graphical relationships among similar word clusters for common skills]]

The heatmap in Figure 4 shows the embeddings of the top 10 most similar words to “Python” in each of the three titles in a color scale. Broadly speaking, across all the dimensions, the most similar words to “Python” appear to have a closer relationship for Data Engineer and Data Analyst than Data Scientist. The color shading across the words for each of the Data Engineer and Data Analyst blocks are less variant than that for the Data Scientist block. Similarly, in Figure 5, we can see that the clusters of top 20 most similar words to “Python” are tighter for Data Engineer and Data Analyst but more spread out for Data Scientist. This may indicate that Data Scientists use Python for a wider set of applications than their Data Engineer and Data Analyst counterparts. In addition, it highlights the domain-specific nature of the Data Scientist role. The ways to use Python for their responsibilities highly depends on the industry and companies these Data Scientists. It would be beneficial that Data Scientist job seekers tailer their skill improvement based on use cases in their targeted industries and companies.

Figure 4. Heatmap on the Top 10 Most Similar Words to “Python” in each Title Category

Background pattern

Description automatically generated

Figure 5. Top 20 Most Similar Word to “Python” in the Transformed t-SNE Space

Chart, scatter chart

Description automatically generated

1. Interview Query’s blog: <https://www.interviewquery.com/p/job-market-update-january-2023> [↑](#footnote-ref-1)
2. Northwestern MSiA blog: https://sites.northwestern.edu/msia/2020/11/30/what-skills-do-data-scientists-need-a-text-analysis-of-job-postings/ [↑](#footnote-ref-2)
3. https://www.betterteam.com/job-boards [↑](#footnote-ref-3)
4. https://chairnerd.seatgeek.com/fuzzywuzzy-fuzzy-string-matching-in-python/ [↑](#footnote-ref-4)