# Skill Demand in the Data Science Job Market

**Group Name, Group Leader, Group Members:** Group MRR, Morgan Simmons, Rachel Kang, Rowan Rosenblum

This project analyzes 2025 data science job postings to evaluate the frequency of requested technical skills and determine whether machine learning and artificial intelligence (ML/AI) related skills dominate employer demand.

**The central research question is:** Which skills are most frequently requested in recent data science job postings, and do machine learning/AI skills rank highest among them?

We hypothesize that ML/AI-oriented skills will rank among the top three most frequently requested skills in data scientist job postings.

## Dataset 1: 2025 Data Science Job Postings (Kaggle)

**Unit of Observation:** Each row in this dataset represents a single job posting for a data science related role.

**1.1 Data Source:** The primary dataset was obtained from Kaggle and consists of publicly available 2025 data science job postings. The dataset includes structured information about each job listing, including job title, company information, industry classification, location, and a list of requested skills.

**1.2 Variables Used:** From the original dataset, the following variables were retained for analysis:

job_title – The posted job title

location – Geographic location of the job

headquarter – Company headquarters location

industry – Industry classification of the company

skills – A string containing listed skills associated with the job posting

Because the skills column contained multiple skills grouped within a single text string, substantial cleaning and restructuring were required prior to analysis.

**1.3 Data Cleaning and Skill Extraction:** The skills variable initially contained list-like formatting (e.g., brackets and quotation marks). These formatting characters were removed using string-processing functions to standardize the text.

After cleaning, the skills column was separated into up to three distinct variables:

Skill_1 Skill_2 Skill_3

Each variable represents one of the top three skills listed in the job posting. If fewer than three skills were listed, remaining fields were recorded as missing (NA). Postings that contained no listed skills were excluded from the analysis. This structured format allowed for consistent counting and comparison of skill frequencies across job postings.

**1.4 Constructed Variables**:To evaluate ML/AI skill dominance, additional variables were created:

Total_Skills – The number of non-missing skill entries among the top three skills.

AI_Skill_Count – The number of extracted skills that match predefined ML/AI-related skills.
AI_Skill_Ratio – The proportion of skills in a posting that are ML/AI-related, calculated as:

$$\text{AI\_Skill\_Ratio} = \text{AI\_Skill\_Count} / \text{Total\_Skills}$$

These constructed variables allow both frequency-based and proportional analysis of ML/AI demand.

# Dataset 2: ML/AI Skills Reference Dataset

**Unit of Observation:** Each row in this dataset represents a single skill classified as ML/AI-related.

This secondary dataset consists of a predefined vector (dictionary) of ML/AI-related skills used for classification purposes. Examples include:

Machine Learning Deep Learning Neural Network PyTorch TensorFlow
Scikit-learn NumPy

Each extracted job skill was cross-referenced against this ML/AI skills list. If a match was identified, the skill was classified as ML/AI-related and included in the AI_Skill_Count variable.

This reference dataset functions as a classification tool rather than as an independent observational dataset.

# Descriptive Statistics

This section presents summary statistics for the cleaned job posting dataset.

| Statistic <chr> | Value <dbl> |
|---|---|
| Total_Postings | 743.000 |
| Avg_Total_Skills | 2.790 |
| Avg_AI_Skills | 0.655 |
| Mean_AI_Skill_Ratio | 0.256 |
| SD_AI_Skill_Ratio | 0.290 |
| Min_AI_Skill_Ratio | 0.000 |
| Max_AI_Skill_Ratio | 1.000 |

Table 1 provides summary statistics for the 743 job postings analyzed. On average, postings list approximately 2.79 skills among the top three extracted skills. The mean number of ML/AI-related skills per posting is 0.655, corresponding to an average AI skill ratio of 0.26. The distribution ranges from postings requiring no ML/AI skills to postings entirely composed of ML/AI-related skills. These statistics indicate that while ML/AI skills are present in the data science job market, they do not overwhelmingly dominate skill requirements on average.

## Skill Frequency Analysis

To evaluate overall demand patterns, the frequency of each skill appearing in the dataset was calculated and ranked.

| Rank <int> | skill <chr> | Frequency <int> | Percentage_of_Postings <dbl> |
|---|---|---|---|
| 1 | python | 375 | 50.47 |
| 2 | r | 306 | 41.18 |
| 3 | machine learning | 219 | 29.48 |
| 4 | sql | 202 | 27.19 |
| 5 | aws | 198 | 26.65 |
| 6 | pytorch | 148 | 19.92 |
| 7 | spark | 147 | 19.78 |
| 8 | pandas | 76 | 10.23 |
| 9 | numpy | 60 | 8.08 |
| 10 | git | 56 | 7.54 |

Table 2 ranks the most frequently requested skills across all job postings. For each skill, the table reports its frequency, the percentage of postings requesting it, and its overall rank. This

ranking allows evaluation of whether ML/AI-related skills occupy the highest demand positions relative to other technical skills such as programming languages and data management tools. The results directly inform the hypothesis regarding the dominance of ML/AI competencies in the data science labor market.

## ML/AI Skill Proportion Analysis

In addition to ranking individual skills, the proportion of ML/AI skills per job posting was analyzed.

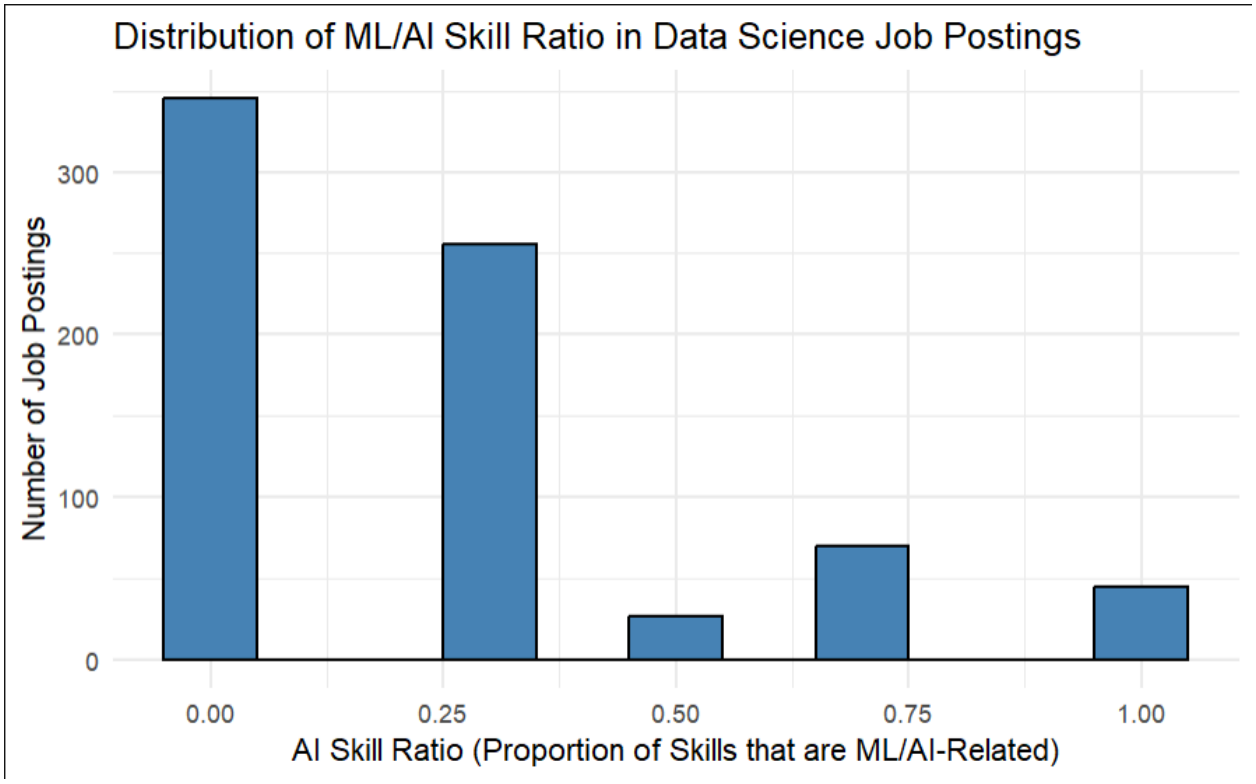


Figure 1 illustrates the distribution of AI skill ratios across 743 job postings. The largest concentration of postings occurs at an AI skill ratio of 0, indicating that many data science roles do not explicitly require ML/AI-related skills among their top listed competencies. A substantial number of postings cluster around a ratio of approximately 0.33, corresponding to one ML/AI skill out of three listed skills. Fewer postings exhibit higher AI intensity, and only a small subset of roles require exclusively ML/AI-related skills. Overall, the distribution suggests that while ML/AI skills are present and relevant, they do not overwhelmingly dominate the broader data science job market.

| AI_Skill_Count <dbl> | Number_of_Postings <int> | Percentage_of_Postings <dbl> |
|---|---|---|
| 0 | 345 | 46.43 |
| 1 | 315 | 42.40 |
| 2 | 77 | 10.36 |
| 3 | 6 | 0.81 |

Table 3 summarizes the number and percentage of job postings requiring zero, one, two, or three ML/AI-related skills among the extracted top skills. This distribution provides insight into whether ML/AI competencies function as central requirements or supplemental qualifications within data science roles. The results suggest that while some roles are heavily ML/AI-focused, many postings either require only one ML/AI skill or none at all, indicating variation in AI intensity across the market.

## Data Dictionary

| job_title | location | headquarter | industry | Skill_1 | Skill_2 | Skill_3 | Total_Skills | AI_Skill_ Count | AI_Skill_ Ratio |
|---|---|---|---|---|---|---|---|---|---|
| Posted job title | Geographic job location | Company headquarters | Industry classification | First listed skill | Second listed skill | Third listed skill | Number of non-missing skills listed | Number of ML/AI-related skills | Proportion of skills that are ML/AI-related |

## Limitations

Several limitations may affect the interpretation of results:

1. **Skill Extraction Constraints** – Skills were limited to the top three listed skills, which may exclude additional relevant competencies mentioned later in job descriptions.

2. **Terminology Variation** – Synonyms, abbreviations, or alternative spellings may lead to undercounting of ML/AI-related skills.

3. **Formatting Irregularities** – Although cleaned, inconsistencies in the original formatting may affect classification accuracy.

4. **Dataset Scope** – The dataset represents postings available through Kaggle's compiled data and may not fully represent the entire U.S. data science labor market.

5. **Binary Classification** – Skills were classified as ML/AI or non-ML/AI, which may oversimplify nuanced skill requirements.

These limitations should be considered when interpreting whether ML/AI skills “dominate” demand.