# DATA ANALYST PROFICIENCY INSIGHT: A COMPREHENSIVE SKILLS AND QUALIFICATIONS ANALYSIS

## 1) MOTIVATION

In today's job market, there's a growing need for people who are good at data analysis. This demand is especially high in sectors like Banking, Media & Entertainment, and Healthcare, where working with big sets of data is common. Many individuals are keen on starting a career in technology, specifically as a data analyst. However, becoming a data analyst involves learning various skills. To make things a bit tricky, different industries look for different skills. For those who are just starting out, figuring out which skills are most important can be a bit confusing.

Our project is here to help with that. We're collecting information from popular job websites and carefully studying job descriptions to figure out the main skills you need to be a successful data analyst. The goal is to make things easier for people who are just starting out, giving them a clear idea of the skills that really matter in the real world of data analysis jobs. This project is all about helping newcomers make smart choices when it comes to developing the skills they need to succeed in the field.

## 2) RESEARCH QUESTIONS

- What skills are most required for Data Analysts?
- What technical skills are most required for Data Analysts?
- What soft skills are most required for Data Analysts?
- What programming languages are most required for Data Analysts?
- What skills are good to have for more opportunities in Data Analyst jobs?

## 3) BACKGROUND AND RELATED WORK

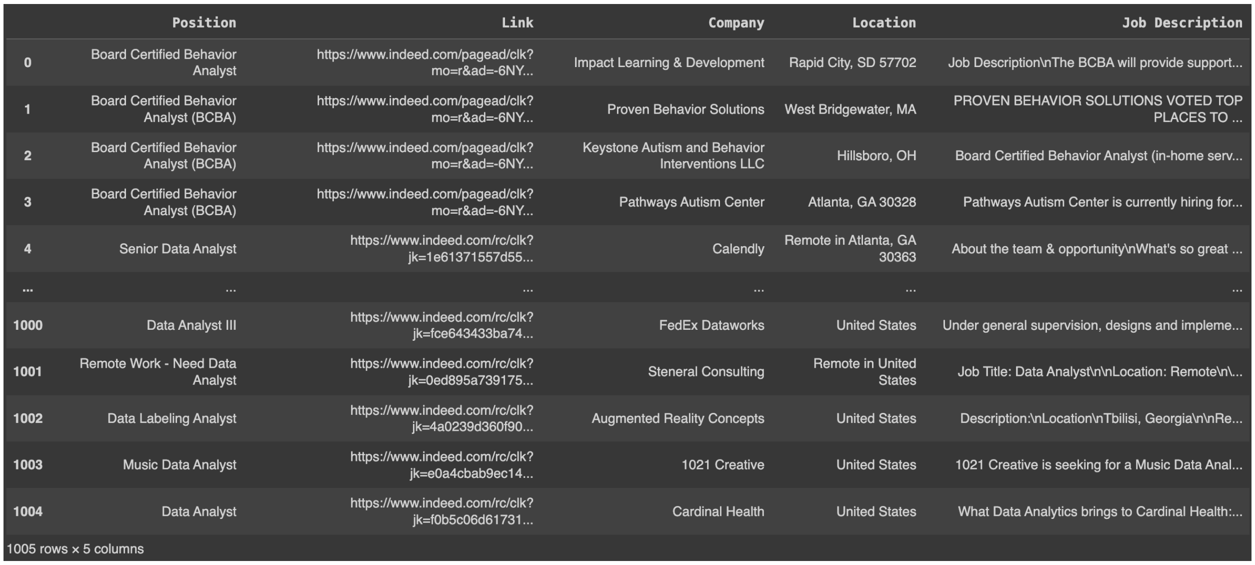
Numerous studies have outlined methodologies for analyzing job descriptions to pinpoint the requisite skills for specific positions. For instance, the research named "Skill Requirements Analysis for Data Analysts Based on Named Entities Recognition," as published by Lina, C. and Jian, Z. (2021), conducted an in-depth exploration into skill requirements for data analysts, utilizing job advertisements from 51job.com. Employing crawlers and manual annotation, the study applied Named Entities Recognition models, with a preference for the Bert pre-training vector model. Simultaneously, another research investigation, "An Investigation of Skill Requirements in Artificial Intelligence and Machine Learning Job Advertisements" by Amit, V. and Kamal, L. (2022), focused on skill prerequisites in AI and machine learning job ads across the USA. Employing content analysis on data from Indeed.com, this study proposed an alternative view of employers' expectations and highlighted crucial skills based on their relative frequency. Both studies contribute valuable perspectives for projects aiming to assist individuals in understanding and cultivating essential skills for success in specific fields, using techniques such as Python programming and Natural Language Processing (NLP). In my project, inspired by these methodologies, I aim to identify required data analyst skills while adding an industry-specific dimension through comparative analysis across different sectors.

## 4) METHODOLOGY AND EVALUATION

The project's methodology employs a multifaceted approach to achieve its objectives. Initially, Python libraries like BeautifulSoup and Scrapy are utilized for web scraping, enabling the extraction of relevant data from leading job websites featuring data analyst positions. Subsequently, the collected job descriptions undergo thorough preprocessing to ensure data integrity and suitability for analysis. This preprocessing includes removing rows with missing values, eliminating stopwords and punctuation, converting text to lowercase, and lemmatizing words to standardize the dataset. Following preprocessing, the dataset is segmented into unigrams and bigrams, facilitating a detailed analysis of individual words and adjacent word pairs. After that, we use both a rule-based approach and a word embeddings approach to match words in the job descriptions with predefined skill keywords, then select the one with better performance for identifying the relevant skills. Finally, the results are visualized to pinpoint which skills are in highest demand in the data analyst job market.

## 4.1 DATA COLLECTION

The dataset is gathered from the Indeed website through dynamic web scraping methods using Python alongside BeautifulSoup and Selenium libraries. The collection process involves automatic scraping of the dataset from the website. To guarantee the acquisition of comprehensive data, the coding is configured to wait for the webpage to fully load and present the data before commencing with the scraping process. For this project, approximately 1,000 Data Analyst job listings are collected for subsequent data analysis. The collected data for each job comprises key attributes such as Position name, Company name, Job link, Location, and Job descriptions, with a primary focus on analyzing the content of the Job descriptions. The collected dataset is stored as DataFrame using Pandas library.



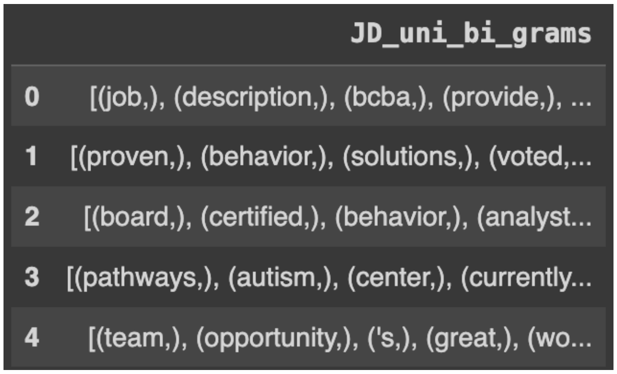
| | Position | Link | Company | Location | Job Description |
|---|---|---|---|---|---|
| 0 | Board Certified Behavior Analyst | https://www.indeed.com/pagead/clk?mo=r&ad=-6NY... | Impact Learning & Development | Rapid City, SD 57702 | Job Description\nThe BCBA will provide support... |
| 1 | Board Certified Behavior Analyst (BCBA) | https://www.indeed.com/pagead/clk?mo=r&ad=-6NY... | Proven Behavior Solutions | West Bridgewater, MA | PROVEN BEHAVIOR SOLUTIONS VOTED TOP PLACES TO ... |
| 2 | Board Certified Behavior Analyst (BCBA) | https://www.indeed.com/pagead/clk?mo=r&ad=-6NY... | Keystone Autism and Behavior Interventions LLC | Hillsboro, OH | Board Certified Behavior Analyst (in-home serv... |
| 3 | Board Certified Behavior Analyst (BCBA) | https://www.indeed.com/pagead/clk?mo=r&ad=-6NY... | Pathways Autism Center | Atlanta, GA 30328 | Pathways Autism Center is currently hiring for... |
| 4 | Senior Data Analyst | https://www.indeed.com/rc/clk?jk=1e61371557d55... | Calendly | Remote in Atlanta, GA 30363 | About the team & opportunity\nWhat's so great ... |
| ... | ... | ... | ... | ... | ... |
| 1000 | Data Analyst III | https://www.indeed.com/rc/clk?jk=fce643433ba74... | FedEx Dataworks | United States | Under general supervision, designs and impleme... |
| 1001 | Remote Work - Need Data Analyst | https://www.indeed.com/rc/clk?jk=0ed895a739175... | Steneral Consulting | Remote in United States | Job Title: Data Analyst\n\nLocation: Remote\n\... |
| 1002 | Data Labeling Analyst | https://www.indeed.com/rc/clk?jk=4a0239d360f90... | Augmented Reality Concepts | United States | Description:\nLocation\nTbilisi, Georgia\n\nRe... |
| 1003 | Music Data Analyst | https://www.indeed.com/rc/clk?jk=e0a4cbab9ec14... | 1021 Creative | United States | 1021 Creative is seeking for a Music Data Anal... |
| 1004 | Data Analyst | https://www.indeed.com/rc/clk?jk=f0b5c06d61731... | Cardinal Health | United States | What Data Analytics brings to Cardinal Health:... |

1005 rows × 5 columns

## 4.2 DATA PREPROCESSING

Before conducting data analysis, it is imperative to convert the collected dataset into a suitable format. Data preprocessing emerges as a critical step in this process, aiming to transform and manipulate the data to render it conducive to subsequent analysis. This step involves several key procedures, including cleaning the dataset by removing rows containing missing values, and tokenizing the text into individual words as both unigrams and bigrams.

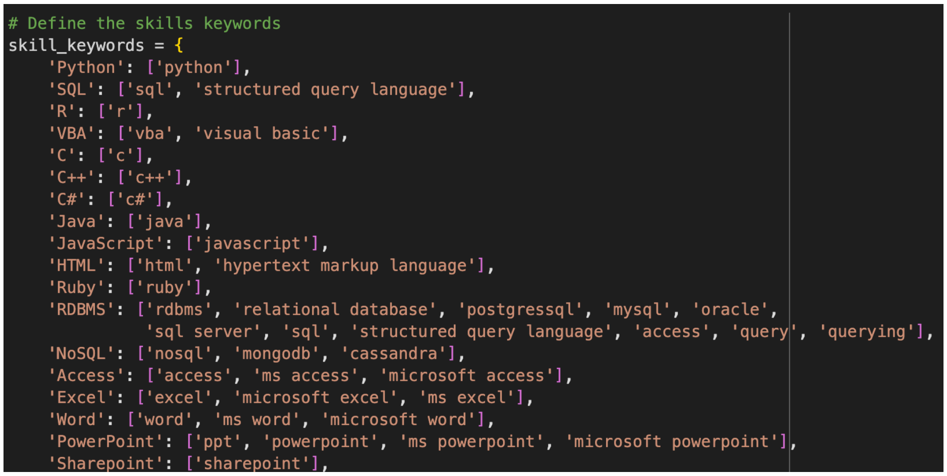
- **Data Cleaning**
  Firstly, the dataset undergoes a review to identify and subsequently remove rows containing missing values. Utilizing Python with the Pandas library, this process is essential to eliminate any insignificant data that may adversely affect the dataset's performance.
- **Tokenization**
  Following data cleaning, the text data is tokenized using the NLTK library in Python. This step aids in identifying and counting the occurrence of required skills for Data Analyst positions by matching individual words within the dataset with predefined skill keywords.
- **Stopword and Punctuation Removal**
  Stopwords and punctuation marks, such as "the," "and," and "is," are then removed from each job description within the dataset using the stopwords and string libraries in Python. This serves to mitigate noise within the text data, ultimately enhancing the accuracy of subsequent data analysis.
- **Lemmatization**
  Subsequently, the tokenized data undergoes lemmatization using the WordNetLemmatizer from the NLTK library. This process reduces tokens to their base form or lemma, thereby reducing the number of unique words within the dataset and contributing to improved analysis accuracy.
- **Unigram and Bigram Conversion**
  To comprehensively cover all words relevant to skills for Data Analyst roles, the dataset is transformed into unigrams and bigrams from the tokenized data. This step accounts for skills that comprise multiple words, such as "Hypothesis testing," "Time Series," and "Project management."



| | JD_uni_bi_grams |
|---|---|
| 0 | [(job,), (description,), (bcba,), (provide,), ... |
| 1 | [(proven,), (behavior,), (solutions,), (voted,... |
| 2 | [(board,), (certified,), (behavior,), (analyst... |
| 3 | [(pathways,), (autism,), (center,), (currently... |
| 4 | [(team,), (opportunity,), ('s,), (great,), (wo... |

## 4.3 FEATURE DEFINITION

A dictionary is constructed containing keywords for each requisite skill. These keywords are then used for matching with individual words derived from the preceding steps, facilitating accurate identification and analysis of essential skills for Data Analyst positions.



```
# Define the skills keywords
skill_keywords = {
    'Python': ['python'],
    'SQL': ['sql', 'structured query language'],
    'R': ['r'],
    'VBA': ['vba', 'visual basic'],
    'C': ['c'],
    'C++': ['c++'],
    'C#': ['c#'],
    'Java': ['java'],
    'JavaScript': ['javascript'],
    'HTML': ['html', 'hypertext markup language'],
    'Ruby': ['ruby'],
    'RDBMS': ['rdbms', 'relational database', 'postgressql', 'mysql', 'oracle',
              'sql server', 'sql', 'structured query language', 'access', 'query', 'querying'],
    'NoSQL': ['nosql', 'mongodb', 'cassandra'],
    'Access': ['access', 'ms access', 'microsoft access'],
    'Excel': ['excel', 'microsoft excel', 'ms excel'],
    'Word': ['word', 'ms word', 'microsoft word'],
    'PowerPoint': ['ppt', 'powerpoint', 'ms powerpoint', 'microsoft powerpoint'],
    'Sharepoint': ['sharepoint'],
```

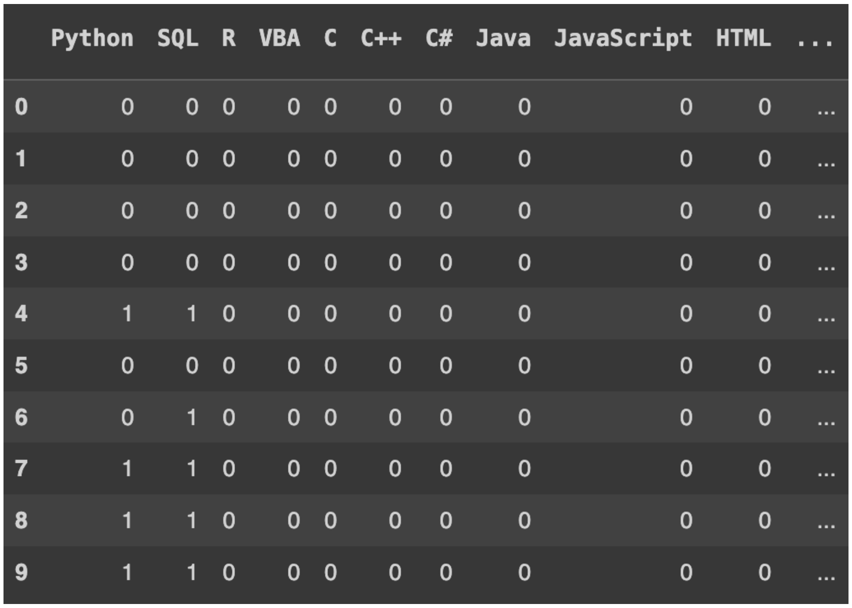
## 4.4 EXPLORATORY DATA ANALYSIS

This step involves counting the occurrence of skills within job descriptions by matching predefined skill keywords with individual words in both unigram and bigram forms. Two approaches are employed: the "Rule-based approach" and the "Word-embeddings approach."

In the Rule-based approach, the Fuzzywuzzy library is utilized to compare the similarity between each skill keyword and individual words. This comparison assesses the similarity between two words character by character.

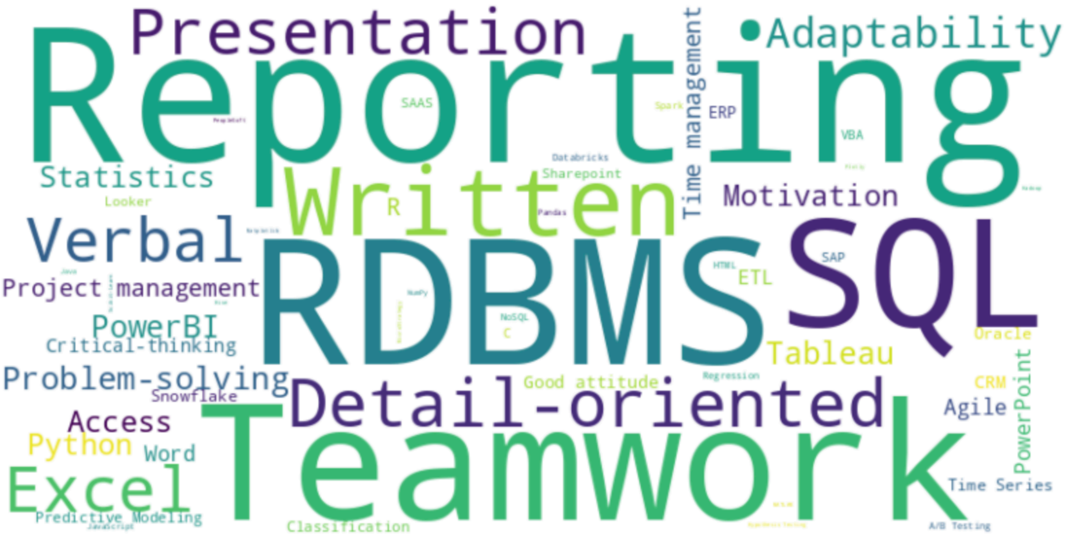
In contrast, the Word-embeddings approach involves embedding words to obtain their respective vectors and then calculating the cosine similarity between these vectors to determine whether they match. If a match is found, the corresponding skill is counted as 1 in the Data Frame, indicating that the job requires that particular skill. Conversely, if no match is found, the skill is counted as 0.

This process allows us to identify the skills most commonly sought after for Data Analyst positions, providing valuable insights into the overall skill requirements within the job market.

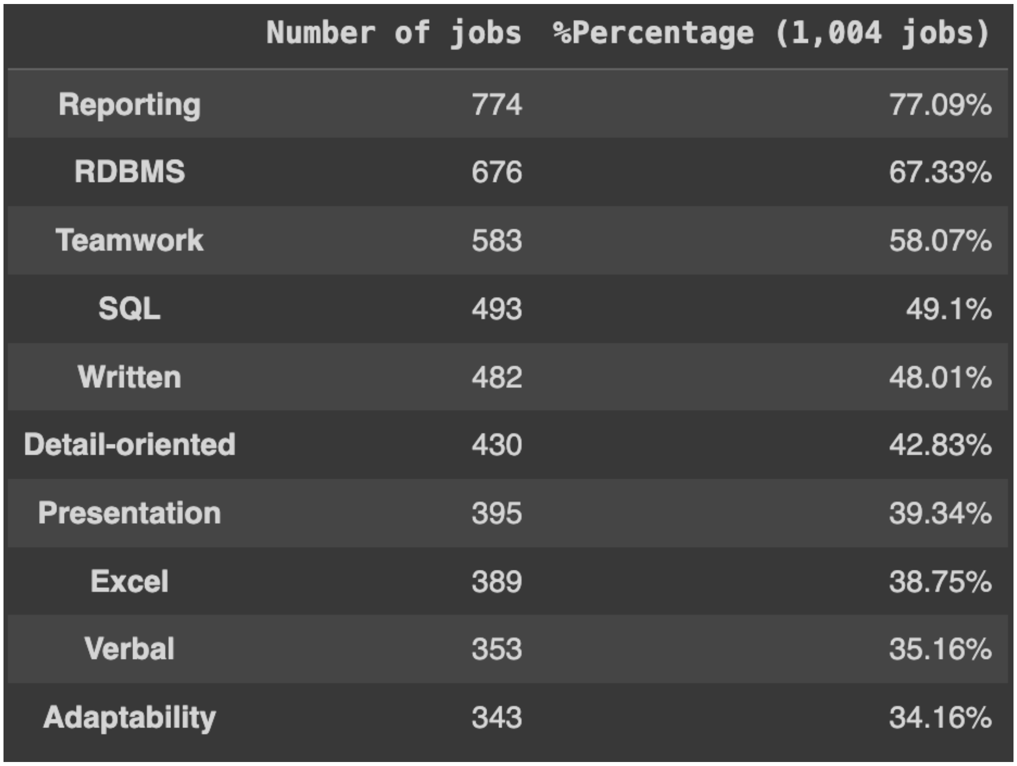


| | Python | SQL | R | VBA | C | C++ | C# | Java | JavaScript | HTML | ... |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... |
| 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... |
| 4 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... |
| 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... |
| 6 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... |
| 7 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... |
| 8 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... |
| 9 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... |

- Word Cloud is generated to visually represent the most in-demand skills for Data Analyst positions. This visualization serves as a valuable tool for identifying prevalent patterns within the dataset, facilitating a deeper understanding of the skill requirements for Data Analyst roles.



- Top 10 skills that are most needed for Data Analyst jobs are visualized in the Table which can also useful for understanding the patterns of the data that Data Analyst jobs tends to require which kind of skills.

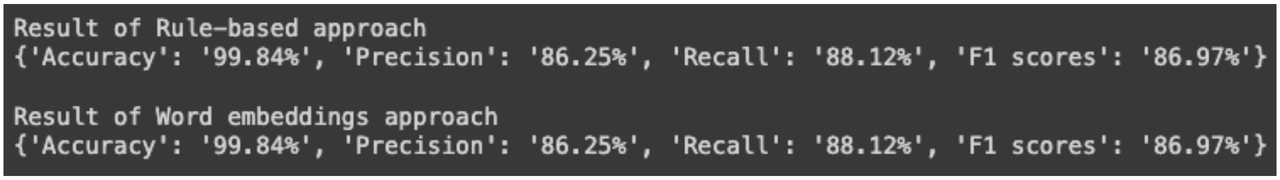


| | Number of jobs | %Percentage (1,004 jobs) |
|---|---|---|
| Reporting | 774 | 77.09% |
| RDBMS | 676 | 67.33% |
| Teamwork | 583 | 58.07% |
| SQL | 493 | 49.1% |
| Written | 482 | 48.01% |
| Detail-oriented | 430 | 42.83% |
| Presentation | 395 | 39.34% |
| Excel | 389 | 38.75% |
| Verbal | 353 | 35.16% |
| Adaptability | 343 | 34.16% |

## 4.5 EVALUATION

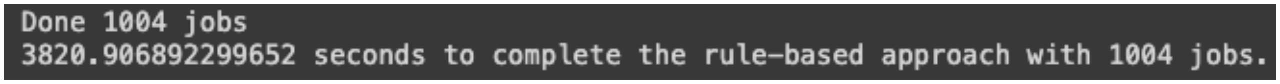
In this project, we have utilized two distinct approaches: the "Rule-based approach" and the "Word-embeddings approach." The aim was to evaluate their respective performances in terms of their ability to accurately detect relevant skills from job descriptions, as well as their computational efficiency. Initially, a manual review of 200 entries from the collected dataset was conducted to extract the requisite skills from each job description. This subset of data served as the test dataset for evaluating the performance of the matching approaches. The evaluation was conducted using metrics such as Accuracy, Precision, Recall, and F1 scores, each providing specific insights into the model's ability to identify skills from job descriptions:

- **Accuracy**
  Accuracy measures the overall correctness of the model's predictions. It indicates the proportion of correctly identified skills compared to the total number of skills in the test dataset.
- **Precision**
  Precision quantifies the accuracy of the model's positive predictions. It measures the proportion of correctly identified relevant skills among all skills predicted by the model as relevant. A higher precision value suggests fewer false positives, indicating a higher degree of confidence in the identified skills.
- **Recall**
  Recall, also known as sensitivity, gauges the model's ability to correctly identify all relevant skills. It measures the proportion of correctly identified relevant skills among all actual relevant skills in the test dataset. A higher recall value indicates fewer false negatives, suggesting that the model effectively captures the relevant skills present in the job descriptions.
- **F1 scores**
  The F1 score is the harmonic mean of precision and recall. It provides a balanced assessment of the model's performance by considering both precision and recall. A higher F1 score indicates a model that achieves high precision and recall simultaneously, thereby offering a comprehensive evaluation of its ability to identify relevant skills from job descriptions.

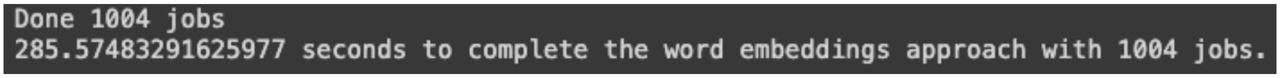


```
Result of Rule-based approach
{'Accuracy': '99.84%', 'Precision': '86.25%', 'Recall': '88.12%', 'F1 scores': '86.97%'}

Result of Word embeddings approach
{'Accuracy': '99.84%', 'Precision': '86.25%', 'Recall': '88.12%', 'F1 scores': '86.97%'}
```



```
Done 1004 jobs
3820.906892299652 seconds to complete the rule-based approach with 1004 jobs.
```



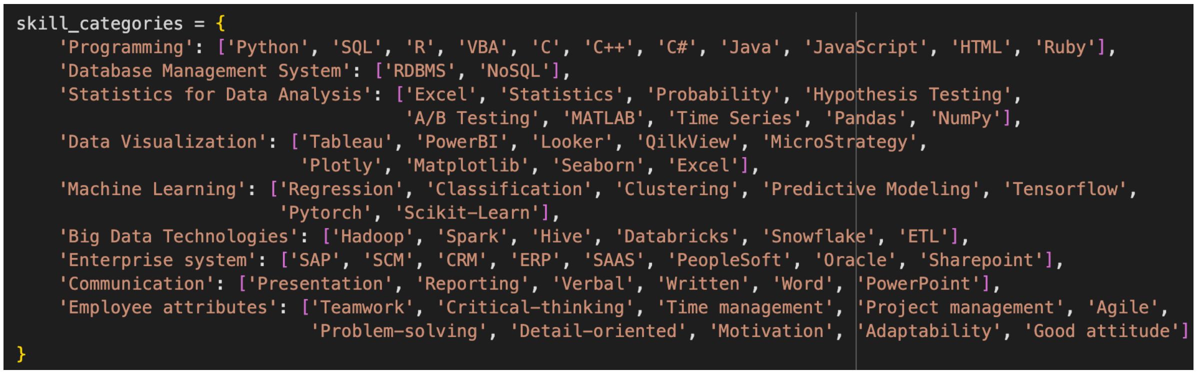
```
Done 1004 jobs
285.57483291625977 seconds to complete the word embeddings approach with 1004 jobs.
```

Both the Rule-based and Word-embedding approaches demonstrated comparable performance, achieving an accuracy of 99.84%, with precision, recall, and F1 score all at approximately 86.25%, 88.12%, and 86.97%, respectively. Notably, the Word-embedding approach completed the task in 285 seconds, significantly faster than the Rule-based approach, which took 3820 seconds. This highlights the Word-embedding approach's efficiency advantage in computational speed while maintaining similar levels of accuracy in identifying skills from job descriptions.

- **Improvement**
  The results indicate that while both approaches demonstrated satisfactory performance, their computational speed remains a concern, particularly when dealing with larger datasets. For instance, with a dataset exceeding 1 million job descriptions, the execution time would likely extend to several hours. To address this issue, an optimization strategy could involve incorporating algorithms to pre-filter words that are unlikely to be relevant skills for Data Analyst positions. By implementing such algorithms, the computational burden on the approaches could be significantly reduced, thereby improving overall execution speed and efficiency.

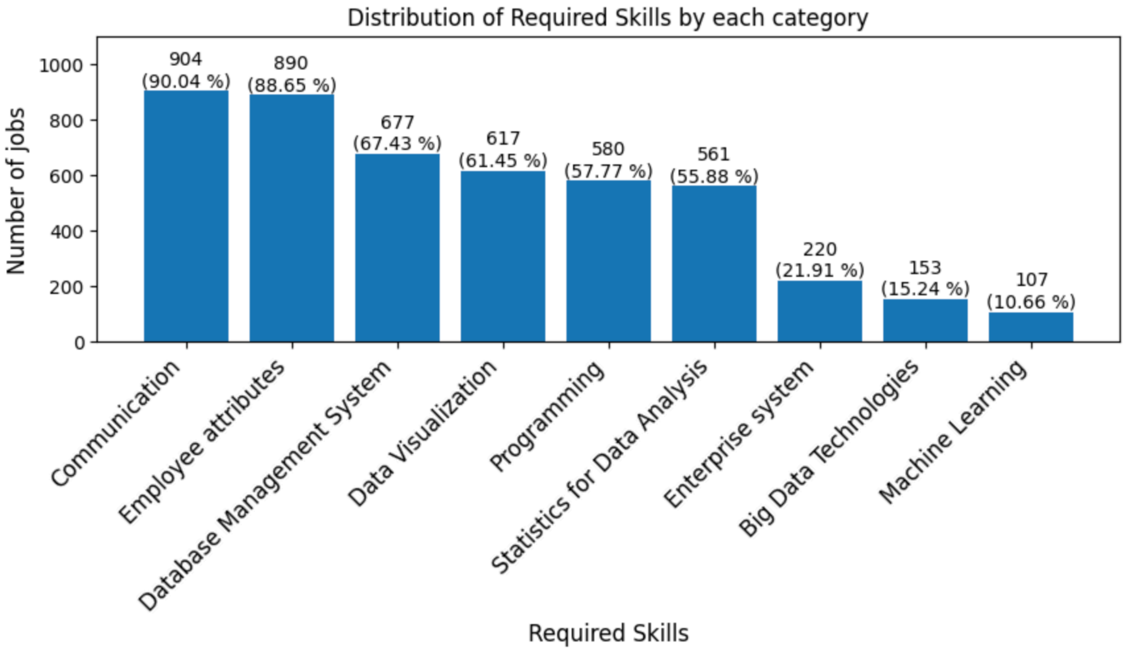
## 5) DATA INTERPRETATION

At the start of the analysis, we organize skills into different groups. This helps us get a clearer picture of how skills are spread out. By dividing skills into categories, we make the data easier to understand, especially for those new to data analysis. This breakdown allows us to see which skills are more important in the Data Analyst job market. This visualization helps people make informed decisions about which skills to focus on. It guides them in developing effective strategies for learning the necessary skills. With a better understanding of which skills are in demand within each category, individuals can tailor their learning efforts to match the needs of Data Analyst roles, improving their chances of success in the job market.

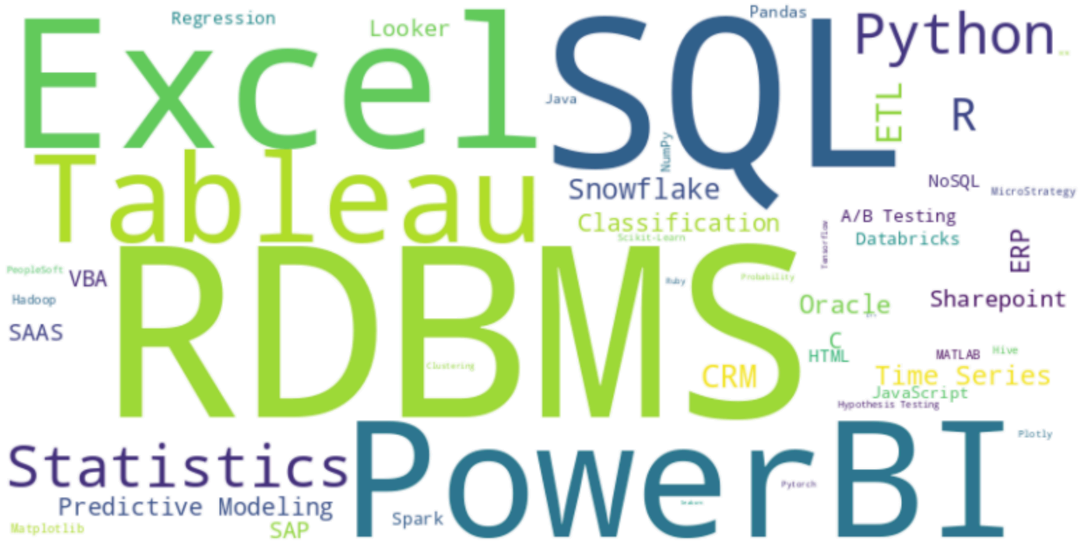


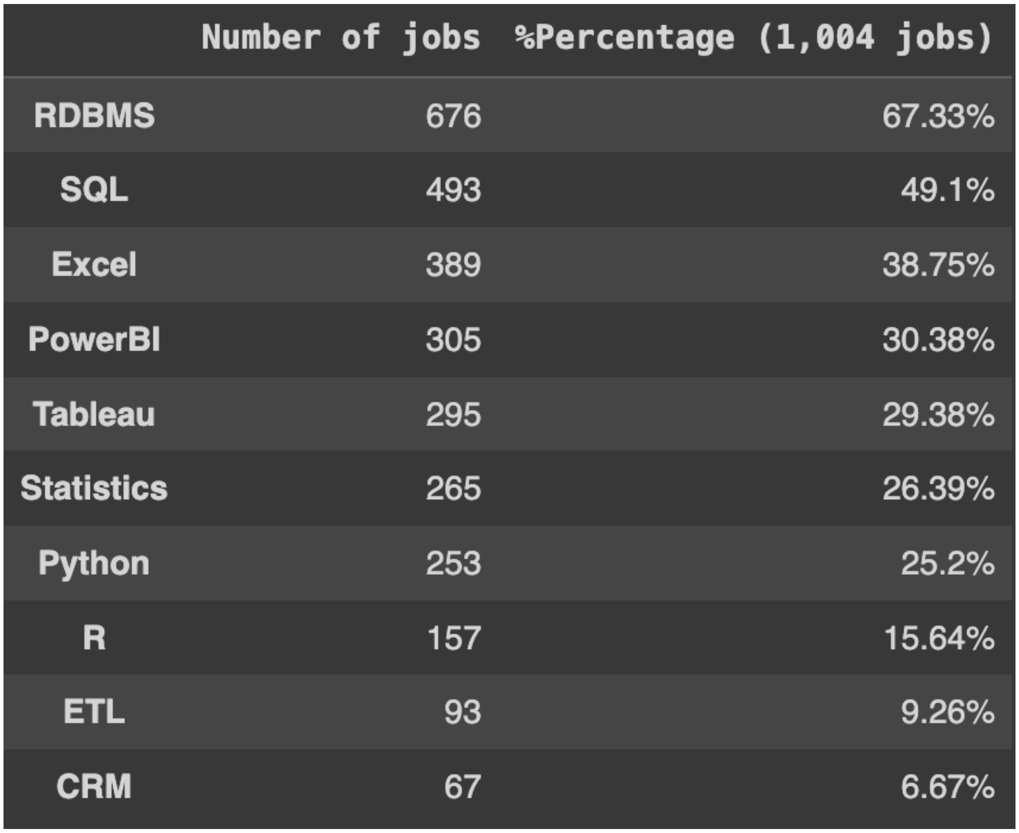
```
skill_categories = {
    'Programming': ['Python', 'SQL', 'R', 'VBA', 'C', 'C++', 'C#', 'Java', 'JavaScript', 'HTML', 'Ruby'],
    'Database Management System': ['RDBMS', 'NoSQL'],
    'Statistics for Data Analysis': ['Excel', 'Statistics', 'Probability', 'Hypothesis Testing',
                                     'A/B Testing', 'MATLAB', 'Time Series', 'Pandas', 'NumPy'],
    'Data Visualization': ['Tableau', 'PowerBI', 'Looker', 'QilkView', 'MicroStrategy',
                           'Plotly', 'Matplotlib', 'Seaborn', 'Excel'],
    'Machine Learning': ['Regression', 'Classification', 'Clustering', 'Predictive Modeling', 'Tensorflow',
                         'Pytorch', 'Scikit-Learn'],
    'Big Data Technologies': ['Hadoop', 'Spark', 'Hive', 'Databricks', 'Snowflake', 'ETL'],
    'Enterprise system': ['SAP', 'SCM', 'CRM', 'ERP', 'SAAS', 'PeopleSoft', 'Oracle', 'Sharepoint'],
    'Communication': ['Presentation', 'Reporting', 'Verbal', 'Written', 'Word', 'PowerPoint'],
    'Employee attributes': ['Teamwork', 'Critical-thinking', 'Time management', 'Project management', 'Agile',
                            'Problem-solving', 'Detail-oriented', 'Motivation', 'Adaptability', 'Good attitude']
}
```

- **Overall Required Skills by categories**
  The process involves tallying the required skills within each predefined category, utilizing the data obtained from the matching approaches. This enables the visualization of the most sought-after skill categories for Data Analyst positions.



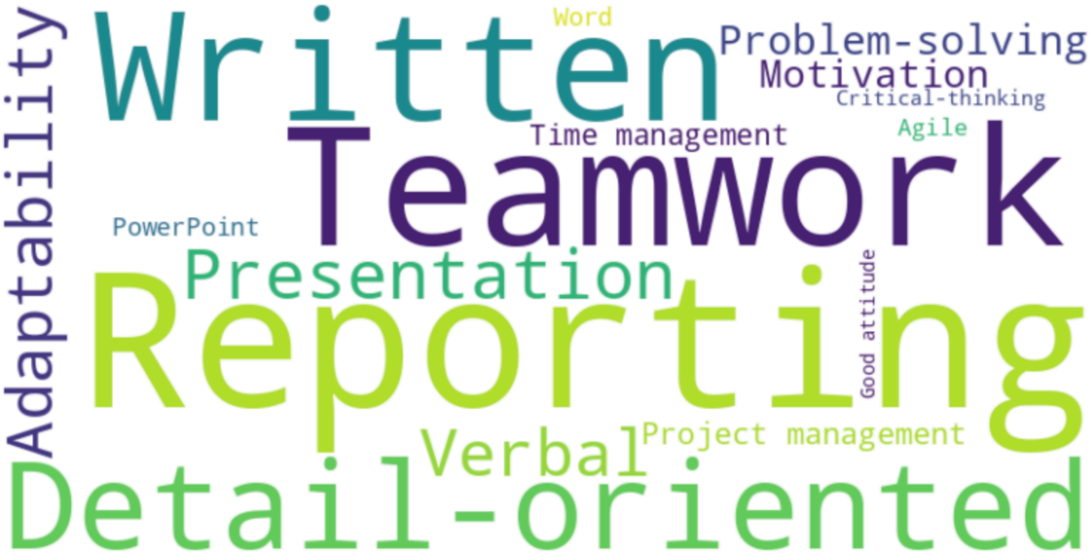
- **Technical Skills**
  To respond to inquiries focused solely on technical skills needed for Data Analyst roles, we've created both a Word Cloud and a Table. These visual aids showcase the most essential technical skills required for such positions.

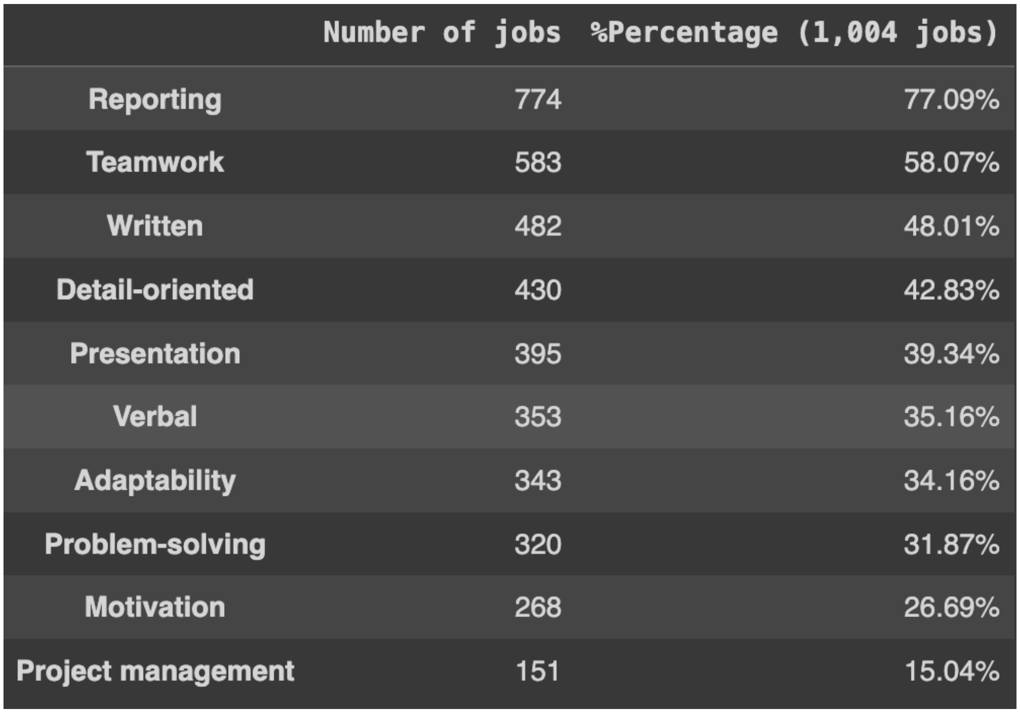




| | Number of jobs | %Percentage (1,004 jobs) |
|---|---|---|
| RDBMS | 676 | 67.33% |
| SQL | 493 | 49.1% |
| Excel | 389 | 38.75% |
| PowerBI | 305 | 30.38% |
| Tableau | 295 | 29.38% |
| Statistics | 265 | 26.39% |
| Python | 253 | 25.2% |
| R | 157 | 15.64% |
| ETL | 93 | 9.26% |
| CRM | 67 | 6.67% |

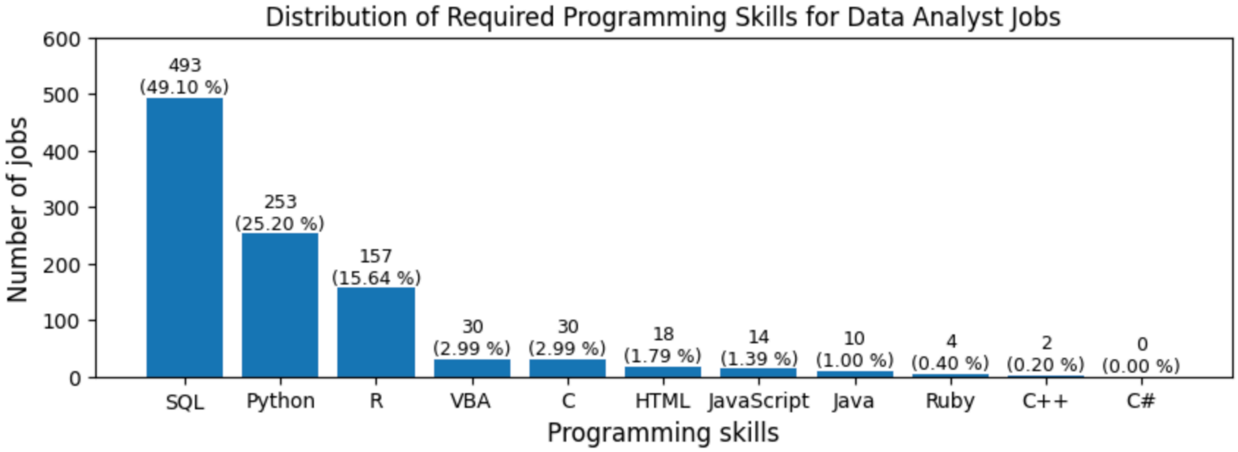
- **Soft Skills**
  To respond to inquiries focused solely on soft skills needed for Data Analyst roles, we've created both a Word Cloud and a Table. These visual aids showcase the most essential soft skills required for such positions.





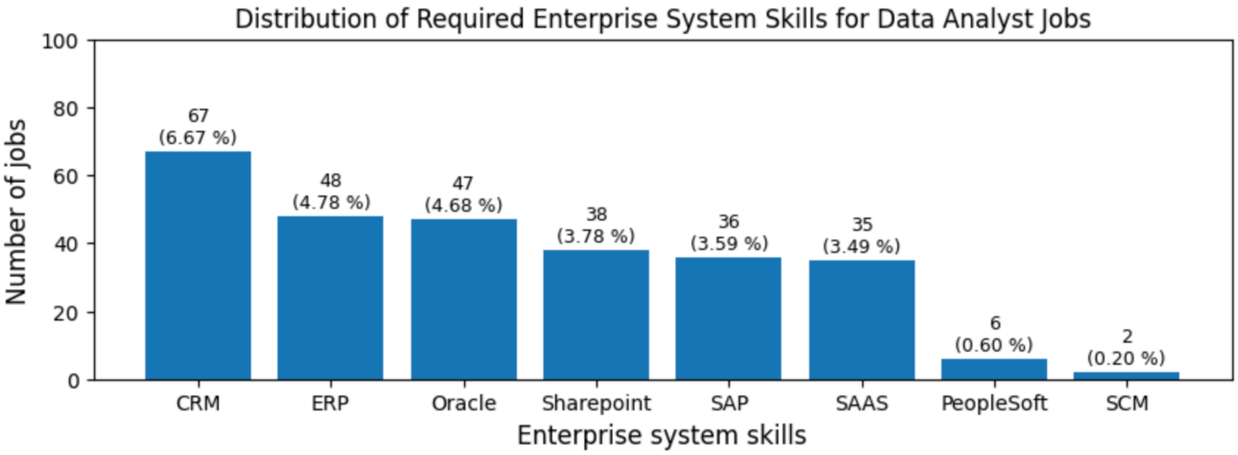
| | Number of jobs | %Percentage (1,004 jobs) |
|---|---|---|
| Reporting | 774 | 77.09% |
| Teamwork | 583 | 58.07% |
| Written | 482 | 48.01% |
| Detail-oriented | 430 | 42.83% |
| Presentation | 395 | 39.34% |
| Verbal | 353 | 35.16% |
| Adaptability | 343 | 34.16% |
| Problem-solving | 320 | 31.87% |
| Motivation | 268 | 26.69% |
| Project management | 151 | 15.04% |

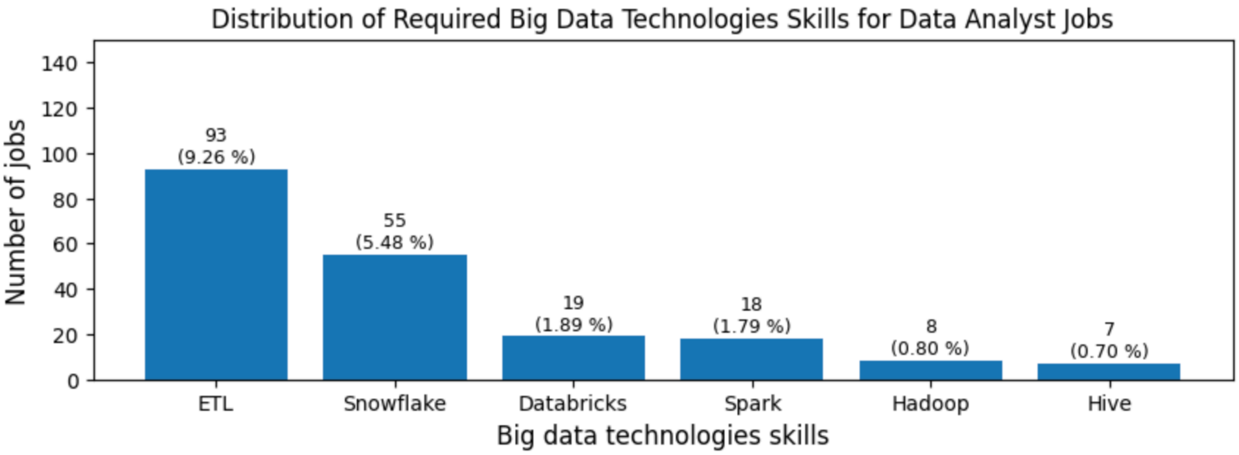
- **Programming Skills**
  To respond to inquiries focused solely on Programming skills needed for Data Analyst roles, we've created a bar chart. These visual aids showcase the most essential Programming skills required for such positions.

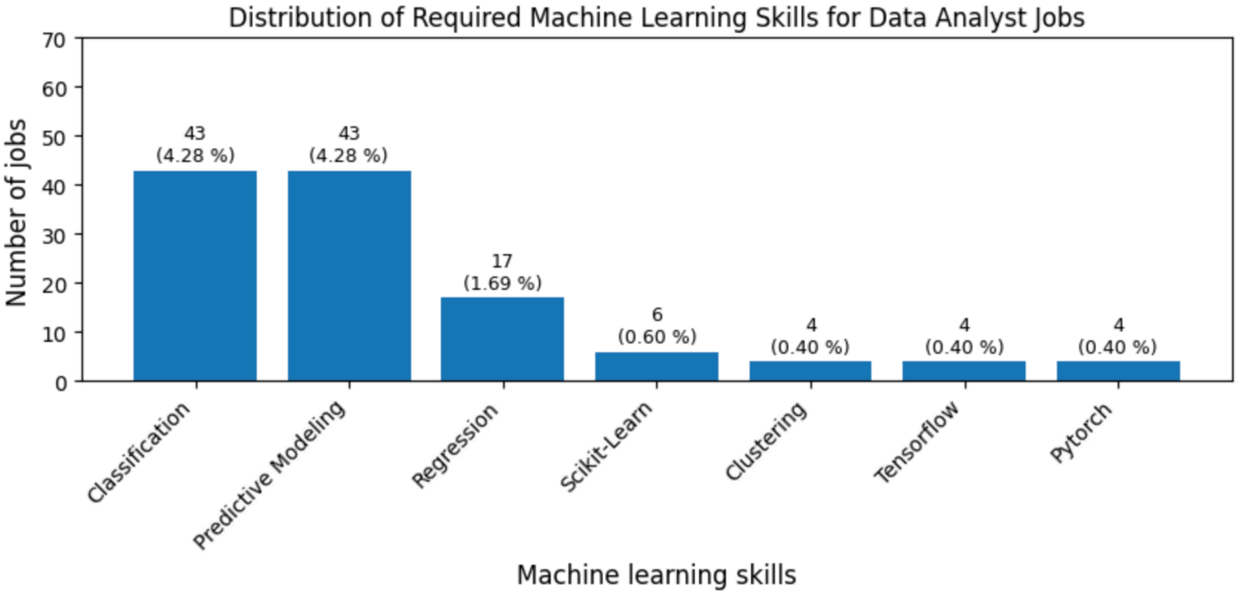


- **Optional Skills**
  Based on the comprehensive analysis of required skills categorized, it's evident that skills related to "Enterprise System," "Big Data Technologies," and "Machine Learning" are least prevalent in Data Analyst job descriptions. This insight is valuable for individuals seeking to broaden their skill set or advance their careers, particularly towards roles such as Data

Scientist. By understanding which skills are less emphasized in Data Analyst positions, individuals can focus their learning efforts on areas that align more closely with their career aspirations.







## 6) CONCLUSION

In summary, this project completed four tasks:

- Data Scraping: We gathered job description data for Data Analyst positions from the Indeed website using Python with BeautifulSoup and Selenium libraries
- Skill Identification Approaches: We implemented both a Rule-based approach and a Word-embeddings approach to identify significant skills required from job descriptions.
- Performance Evaluation: We assessed the performance of both the Rule-based and Word-embeddings approaches in terms of their ability to identify skills from job descriptions and their computational speed. This evaluation was conducted using various metrics, including Accuracy, Precision, Recall, and F1 scores.
- Skill Visualization: We visualized the identified skills to highlight the top skills needed for Data Analyst jobs across different categories, including overall top skills, top technical skills, top soft skills, top programming skills, and top optional skills.

In conclusion, our project serves as a valuable resource for individuals seeking to enter the Data Analyst field by providing insights into the most important skills required for such positions. Additionally, we have shed light on the performance of different skill identification approaches, namely the Rule-based and Word-embeddings methods, offering valuable guidance for skill development and decision-making in this domain.

## 7) FUTURE WORK

Future work may involve expanding the study to analyze a larger pool of job descriptions and identify advanced skills, as well as addressing a wider range of inquiries. This expansion could include examining skills needed for various job roles beyond Data Analyst positions, providing insights into broader job markets. Additionally, future work may involve exploring alternative techniques to enhance the performance of the project. This could include investigating additional methods or approaches that could improve the accuracy and efficiency of skill identification and analysis, thereby providing more robust insights into job market trends and skill requirements.