**Introduction**:

In recent years, rising unemployment has heightened the need to analyze job market trends and gain insights into the evolving landscape. Prominent job portals like Indeed, Naukri, and Glassdoor have shown a notable increase in demand for positions such as Data Scientist, Machine Learning Engineer, and Data Engineer. To effectively track these trends, web scraping has become an essential technique. Python, with its powerful libraries and tools, has played a pivotal role in extracting and analyzing data

For data science and machine learning positions in India, In this project, data was gathered through **Web scraping** from popular job portals like **Indeed.com**. The primary objective is to uncover current

Job market trends, identify key skills in demand, and determine the ideal job positions for data science professionals.

### **2. Data Collection Method**

#### ****Sample Size & Process of Data Collection****

The data was collected from Indeed.com, focusing on job postings for positions such as Data Scientist, Data Engineer, and Machine Learning (ML) Engineer. The keywords used for the search were “Data Scientist,” “Data Engineer,” and “ML Engineer.”

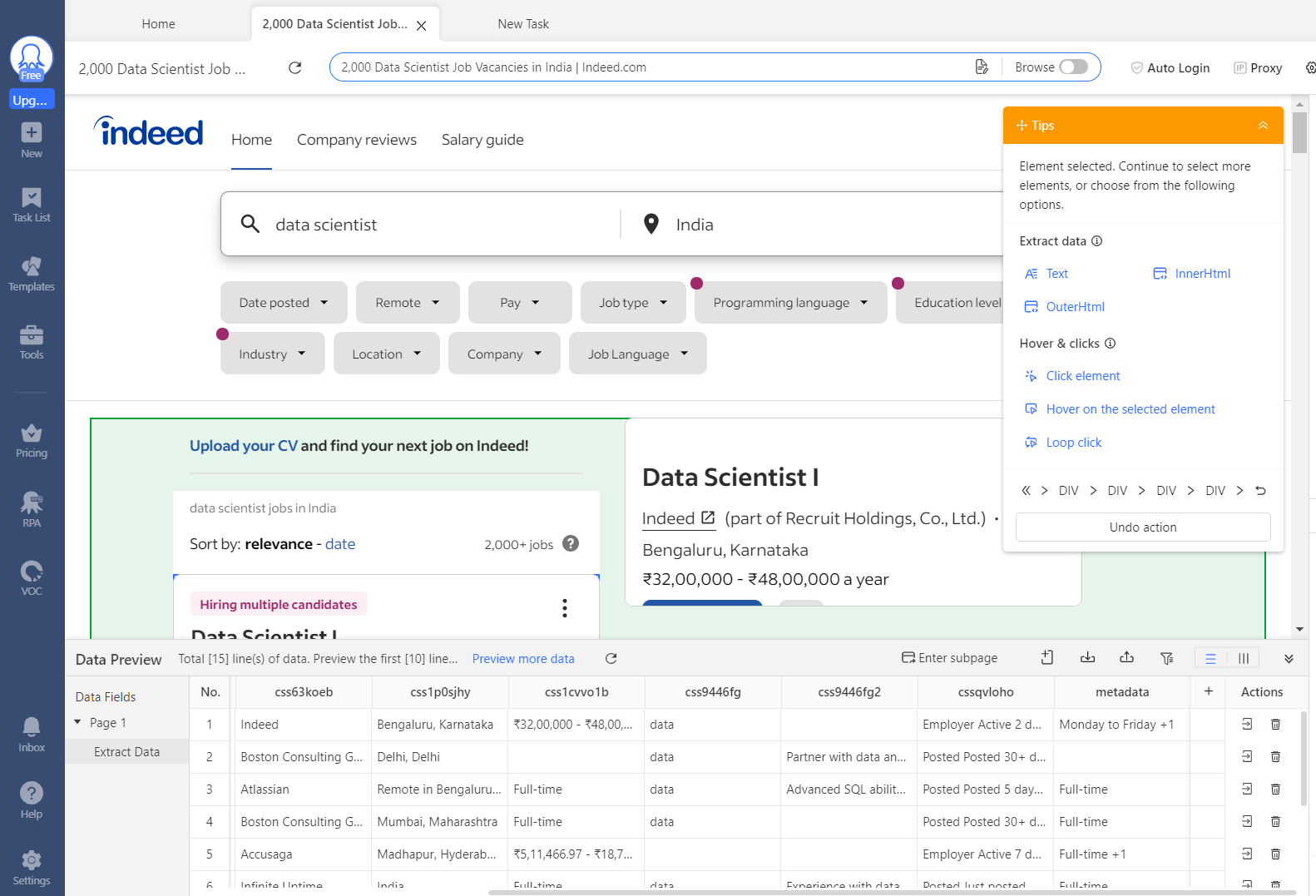
#### ****Data Collection Process****

The initial plan involved using Python libraries like BeautifulSoup and Selenium for web scraping. However, these methods encountered technical difficulties, particularly ProxyError issues, due to unstable network connections and restricted access to the website. This hindered the extraction process. In this case, persistent ProxyError issues, likely due to blocked or unreliable proxies, made these tools impractical for the task.

#### ****Switching to Octoparse****

To overcome these challenges, Octoparse, a free web scraping desktop application, was employed. Octoparse effectively handled the data extraction process without encountering the issues faced with Python libraries. Data was successfully scraped from Indeed.com and stored in the DataScientistJobVacancies\_Indeed.com.csv file.

**OctoParse Application**



#### ****Data Extraction:****

#### **The following data was extracted from Indeed website:**

#### **Job Position**

#### **Company**

#### **Location**

#### **Job Type**

#### **Posted Date**

#### **Key Skills**

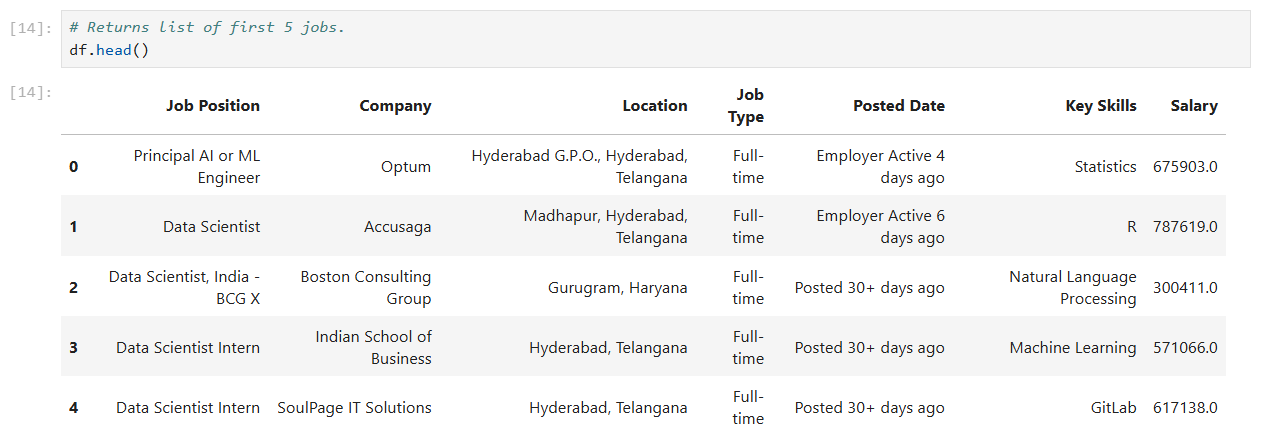
#### **Salary**

#### ****Data Cleaning and Visualization****

Post-extraction, the data was cleaned by removing duplicates, handling missing values, and perform exploratory data analysis using Pandas. The initial five job positions were inspected to ensure data integrity.

The following analyses were performed:

* **Job Position Frequency**: The frequency of different job positions, such as Data Scientist, Data Engineer, and Machine Learning Engineer, was calculated.
* **Visualizations**: Using Python’s Matplotlib and Seaborn libraries, various visualizations were created:
  + A Word Cloud to visualize the most common job positions.
  + Charts showing the distribution of job locations (remote vs. onsite).
  + Analysis of Full-Time vs. Part-Time positions.
  + Calculation and visualization of the average salary for each job position.



Salaries were extracted and converted into Number data type in excel. Key skills were extracted and converted into Text data type.

**2. Market Data Visualization:**

a. Distribution of Job Titles:

The provided bar chart illustrates the frequency of the top ten job positions related to data science. These positions are ranked by the number of job postings in a specific dataset or market. Let’s delve into the key trends:

The most frequently advertised job position is “Data Scientist,” with well over 200 listings.

This high demand suggests that organizations are actively seeking professionals with expertise in data analysis, machine learning, and statistical modeling.

Machine Learning Engineers Follow Closely:

The second most common role is “Machine Learning Engineer,” appearing approximately half as often as “Data Scientist.”

Machine learning expertise is highly sought after, although it is not as ubiquitous as general data science skills.

**Specialized Roles:**

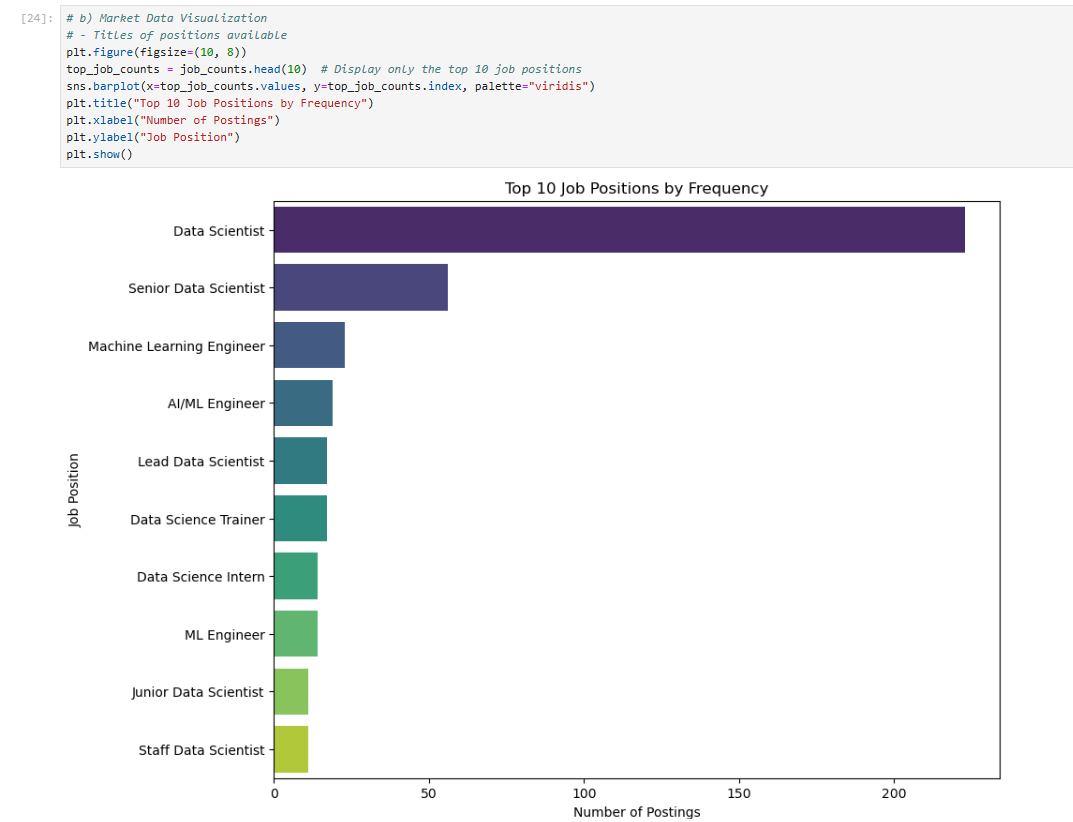
Positions such as “Senior Data Scientist,” “ML Engineer,” and “Data Science Consultant” show gradual decreases in frequency (from around 100 to fewer than 50 postings).

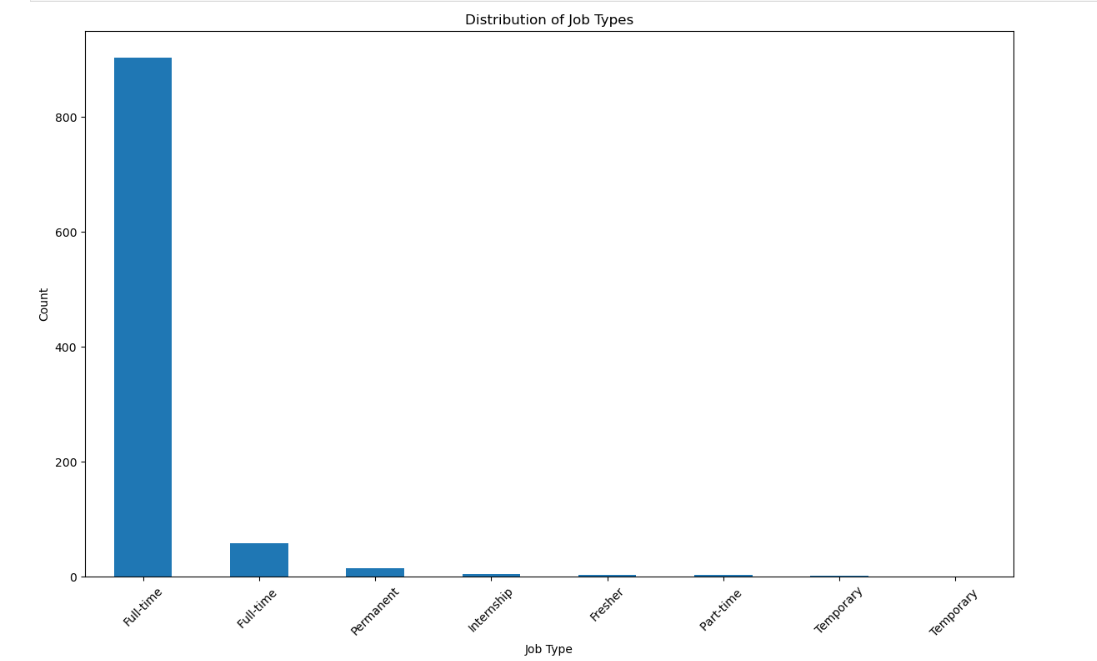
These roles likely require more specialized skills or experience compared to entry-level positions.

**Senior and Entry-Level Positions:**

“Lead Data Scientist,” “AI/ML Engineer,” and “Director of Data Science” are less frequently advertised but likely represent senior or specialized roles.

Interestingly, both entry-level (“Junior Data Scientist”) and highly experienced positions (“Staff Data Scientist”) appear at similar frequencies (fewer than 25 postings each).





**Distribution of Job Types Across India:**

The most prevalent job type is “Full-time” with a count exceeding 800.

Full-time roles seem to dominate the dataset, indicating a significant demand for employees coming to office. It also indicates that the **Organizations** are actively hiring full-time employees.

“Internship” and “Part-time” positions have minimal representation, suggesting limited availability.

These roles may be less common or specific to certain industries or organizations.

