Extraction and Analysis of Job Listings

(Web-Scraping and EDA)

Shivam Adbhute

shivamam26@gmail.com
Indian Institute of Technology Guwahati
GITHUB

ABSTRACT

This project aims to create an intelligent tool to enhance data science job searches by employing web scraping techniques on the TimesJobs website. The tool extracts crucial details from job listings, including job titles, company names, required skills, posting times, locations, and salaries. Through a combination of web scraping, data cleaning, and exploratory data analysis (EDA), the project provides valuable insights into the current data science job market. The tool is designed to assist professionals, job seekers, and recruiters in making informed decisions based on industry trends.

PROBLEM STATEMENT

Develop a tool using web scraping and data analysis to navigate the complexities of the data science job market. This tool should gather, analyze, and predict trends within the data science job market, empowering professionals, and recruiters.

DATASET

1. **Job Title**: The specific designation associated with the job opening.
2. **Company:** The name of the organization that has posted the job.
3. **Skills Required:** The essential skills and qualifications needed for the job.
4. **Job Posted Ago:** The number of days elapsed since the job was posted, providing insight into its freshness.
5. **Location:** The list of cities where the job opportunity is available.
6. **Salary (Lacs p.a.):** The salary range for the position on an annual basis, denoted in lakhs.
7. **Experience Required (Years):** The number of years of professional experience required for the job.

WEB SCRAPING and its application in the project-

This project utilizes web scraping, also called web harvesting or web data extraction, to automate data collection from the TimesJobs website. This allows us to gather information about data science jobs without manual page analysis.

Web scraping process:

1. **Fetching:** Downloading the website's HTML code, like how browsers work
2. **Parsing:** Identifying the desired data elements within the downloaded code.
3. **Extracting**: Isolating and storing the target data in a suitable format.

Benefits of web scraping:

1. **Efficient data collection:** Scraping gathers large amounts of data quickly, saving time and effort.
2. **Targeted data extraction:** Scrapers focus on specific data, improving efficiency.
3. **Trend and pattern analysis:** Analyzing collected data reveals valuable insights.
4. **Task automation:** Scraping automates repetitive tasks like price monitoring and competitor analysis.

Ethical considerations:

1. Respecting robots.txt files and avoiding prohibited websites.
2. Avoiding overloading websites to prevent crashes.
3. Using scraping responsibly and ethically.

Project application of web scraping:

1. **Function Definitions:**

   - **`extract_salary`**: Extracts salary information from job listings by removing unnecessary characters and formatting the data for consistency.

   - **`scrape_jobs:`** Uses BeautifulSoup to iterate through TimesJobs pages, scraping details such as job titles, companies, skills, locations, and salaries.

2. **Data Scraping:**

   - The scrape_jobs function gathers data from the first 10 pages of data science job listings on TimesJobs.

   - It sends HTTP requests to the website, parses HTML content, and extracts relevant information using BeautifulSoup.

3. **Specific Data Extraction:**

   - extract_salary focuses on identifying salary information, extracting elements with "Lacs", and formatting the data appropriately.

   - Experience data is retrieved by locating elements with "yrs" and cleaning the extracted text.

4. **Dataframe Creation:**

   - The extracted data is compiled into a single dataframe that includes job title, company, skills, location, salary, and experience.

   - This dataframe facilitates further analysis and visualization of the collected job data.

Data Cleaning and Preprocessing

Cleaning and preprocessing data are critical stages in refining datasets for analysis. The data cleaning process entails identifying and rectifying errors, addressing missing or inaccurate data, and ensuring the overall quality of the dataset. Its primary goal is to improve the reliability of the dataset.

The datasets underwent several specific actions to ensure they were analysis-ready:

Show Dataset Rows & Columns count Before Removing Duplicates:

Rows count: 250 Columns count: 7

Remove duplicates: `df.drop_duplicates(inplace=True)`

Show Dataset Rows & Columns count After Removing Duplicates:

Rows count: 248 Columns count: 7

Show Dataset Rows & Columns count Before Removing Missing Values:

Rows count: 248 Columns count: 7

Replace empty strings with NaN in the 'Location' column:

```
df['Location'].replace('', pd.NA, inplace=True)
```

Drop rows with null values in the 'Location' column:

```
df.dropna(subset=['Location'], inplace=True)
```

Show Dataset Rows & Columns count After Removing Missing Values:

Rows count: 236 Columns count: 7

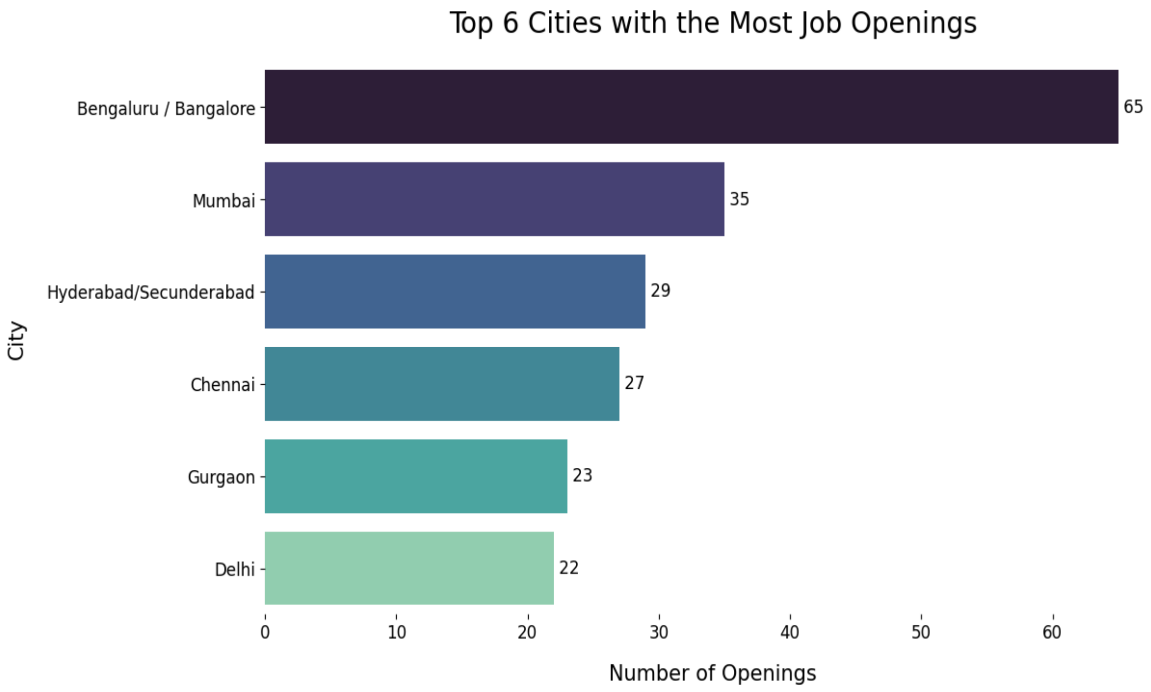
Check missing values again to confirm: `df.isnull().sum()`

Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a critical phase in the data analysis journey. It involves a comprehensive examination and analysis of a dataset to grasp its inherent characteristics, highlight key features, and uncover potential patterns or relationships within the data.

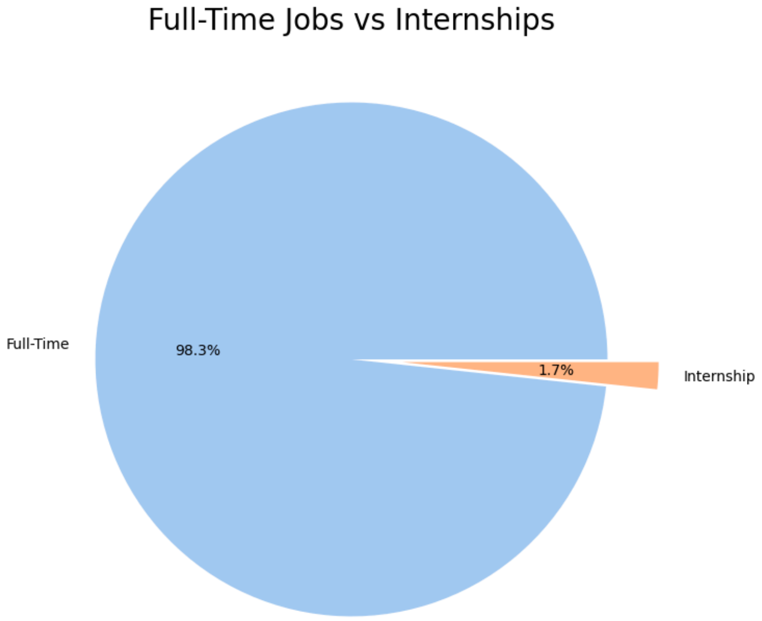
Top 6 Cities with the Most Job Openings in Data Science:

Bar charts effectively represent the frequency of occurrences across different levels of a categorical variable. In this instance, a bar chart is employed to identify the top 6 cities with the highest job openings in Data Science. The visualization illustrates Bengaluru/Bangalore as the leader in job openings, trailed by Pune, Chennai, and Mumbai, with Ahmedabad and Gurgaon appearing at the lower end of the list.

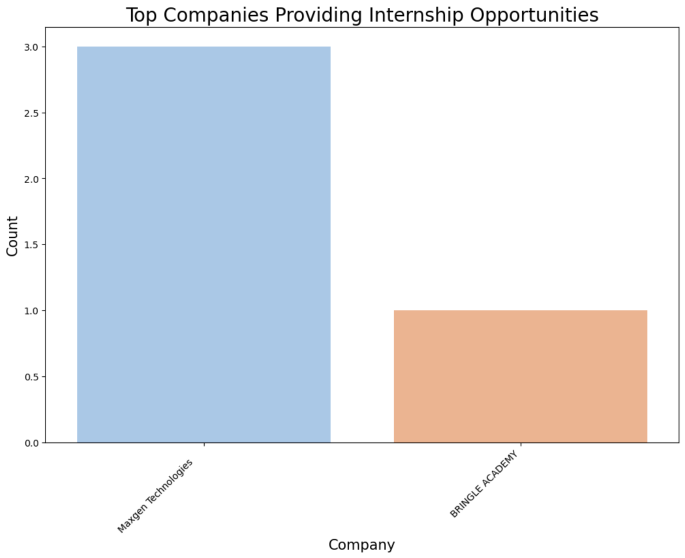


Comparison of Full-Time Jobs and Internships in the Job Market:

Here, the pie chart indicates that approximately 97.2% of data science job opportunities are full-time positions, emphasizing the substantial demand for permanent roles. Conversely, internships constitute a modest 2.8% of the opportunities.

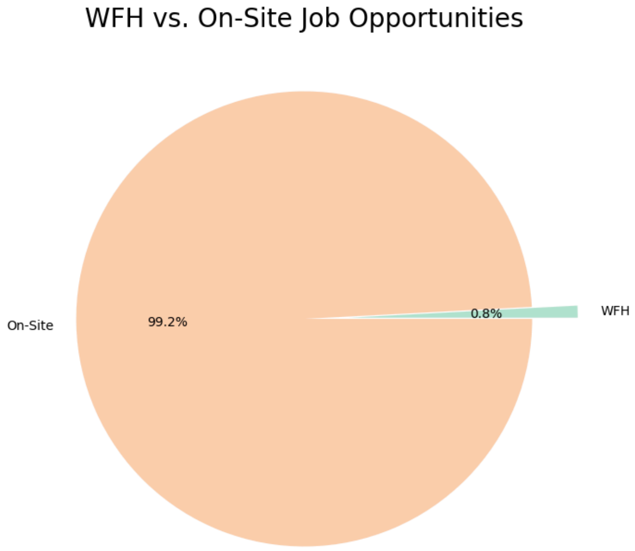


Top Companies Providing Internship Opportunities in the Job Market:

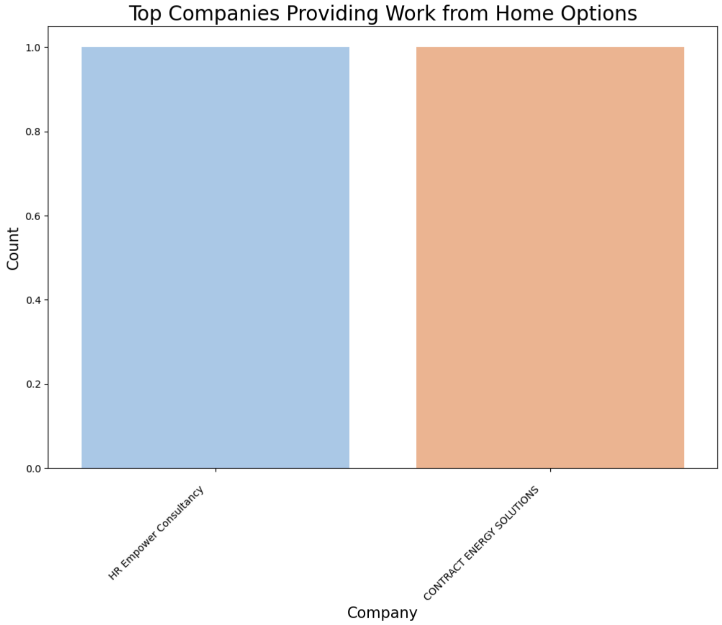


Comparison of Work from Home vs On Site Job Opportunities in the Job Market:

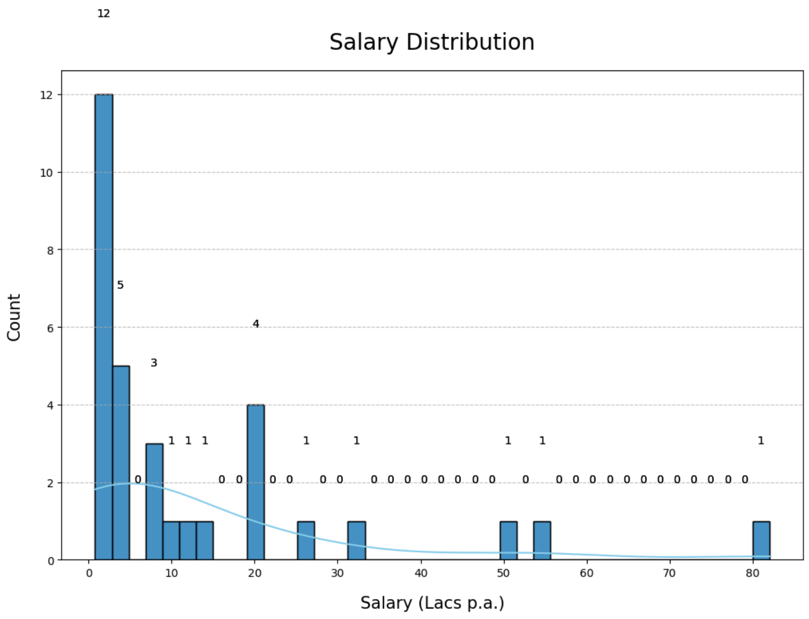
The pie chart communicates the current proportions of work-from-home and on-site opportunities in the job market, revealing that 96.4% of jobs require on-site presence, while 3.6% offer work-from-home options.



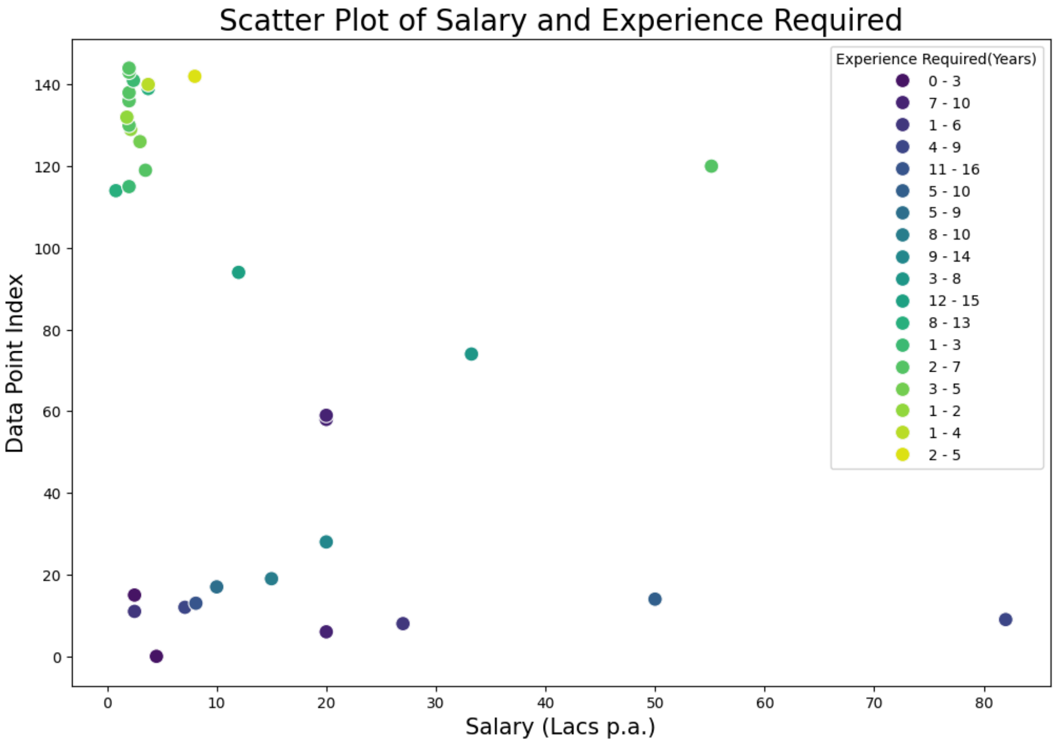
Top Companies Providing Work from Home Options in the Job Market:



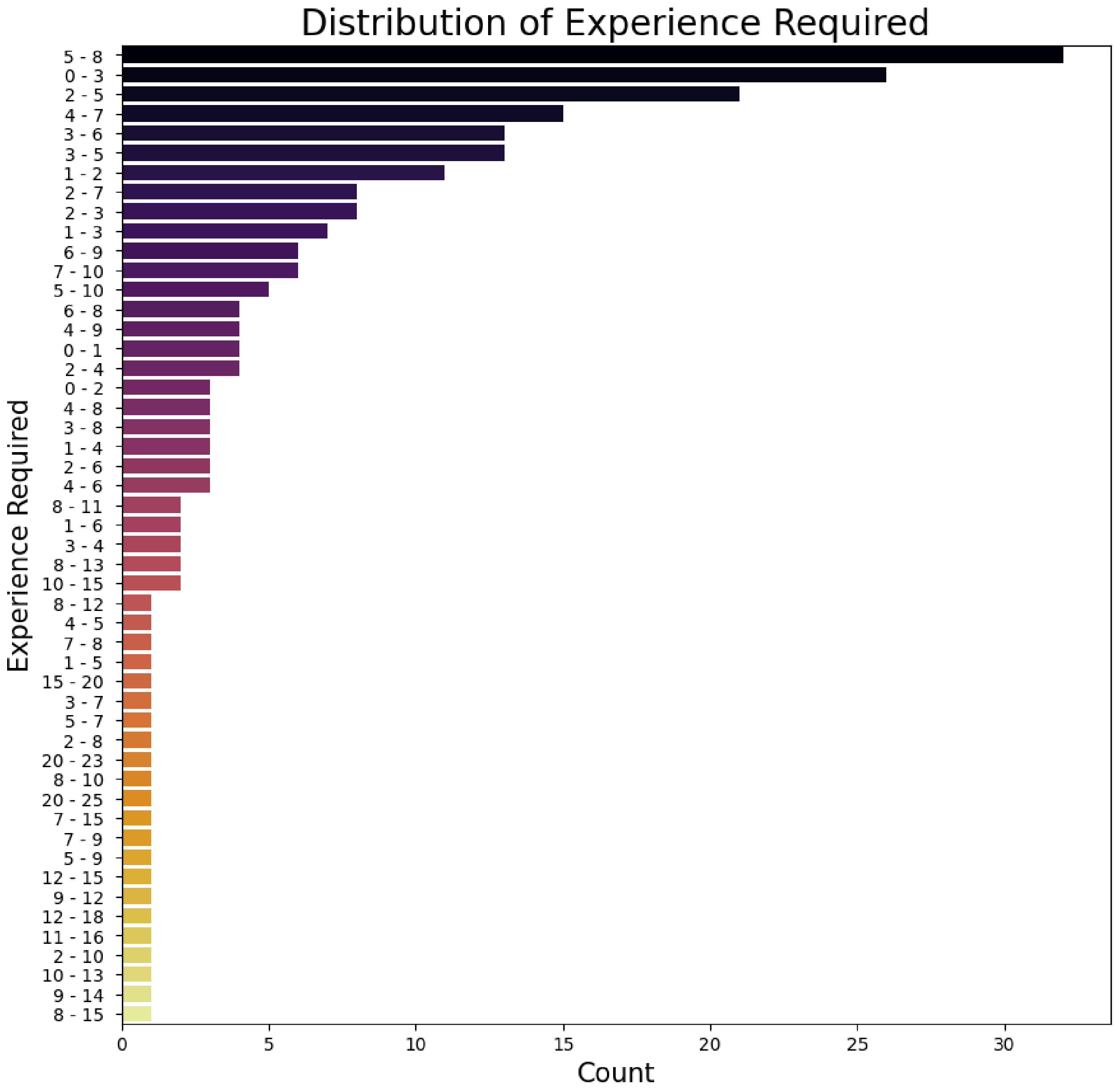
Salary Distribution in the Job Market:



Analyzing the Relationship Between Salary and Experience in the Job Market:



Experience Requirements in the Job Market:



Conclusion

The analysis of data scraped from TimesJobs reveals several key insights into the data science job market in India. Python, SQL, Machine Learning, and Data Analysis are the most sought-after skills, with Bengaluru leading in job openings. Full-time positions dominate the market, while on-site presence is more common than remote work. Entry-level salaries are clustered around 0-10 Lacs per annum, while experienced professionals can expect significantly higher packages. A notable demand exists for individuals with varying levels of experience, ranging from entry-level to seasoned professionals. This analysis further reveals a moderate positive correlation between average salary and average experience, indicating that salaries generally increase with experience. This project successfully developed an intelligent tool that enhances data science job search efficiency through web scraping. The tool leverages data analysis and visualization techniques to provide valuable insights into the job market, serving as a valuable resource for professionals, job seekers, and recruiters. However, it's important to note that the insights captured represent a snapshot of the dynamic market and may evolve over time. Nevertheless, the project contributes significantly to enhancing accessibility and informed decision-making within the ever-changing landscape of data science employment.