CAPSTONE PROJECT

WEB SCRAPING DATA SCIENCE JOB LISTINGS

BY

Sachin Verma

EMAIL – vermasachin1505@gmail.com

GITHUB LINK

https://github.com/sachin798582/Web_Scraping_Data_Science_Job_Listings

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.

Keywords: Web Scraping, Exploratory Data Analysis (EDA), In-Demand Skills, Top Companies, Full-Time Jobs, Internships, Salary Distribution

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, analyse, and predict trends within the data science job market, empowering professionals, and recruiters.

Business Objective

Analyse data science job market trends through web scraping and data analysis.

Datasets Overview

Descriptions for Scraped Dataset

Job Title: The specific designation associated with the job opening.

Company: The name of the organization that has posted the job.

Skills Required: The essential skills and qualifications needed for the job.

Job Posted Ago: The number of days elapsed since the job was posted, providing insight into its freshness.

Location: The list of cities where the job opportunity is available.

Salary (Lacs p.a.): The salary range for the position on an annual basis, denoted in lakhs.

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:

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

Applications of web scraping:

- Price comparison: Websites like Google Shopping compare prices from various retailers.
- Market research: Companies use scraping to gather competitor, target market, and industry data.
- Social media analysis: Businesses track brand mentions and sentiment through scraping.
- Dataset building: Researchers use scraping to build large datasets for various purposes.

Ethical considerations:

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

Project application of web scraping:

- Function definition:
  - `extract_salary`: Extracts salary information from job listings, removing unnecessary characters and formatting the data.
  - `scrape_jobs`: Iterates through TimesJobs pages, scraping data like job title, company, skills, location, and salary using BeautifulSoup.
- Data scraping:
  - `scrape_jobs` extracts data from the first 10 pages of data science job listings on TimesJobs.
  - Requests are sent to the website, HTML content is parsed, and desired information is extracted using BeautifulSoup.
- Specific data extraction:
  - `extract_salary` focuses on extracting salary information, identifying elements containing "Lacs" and formatting the data.
  - Experience information is extracted by searching for elements containing "yrs" and cleaning the text.
- Dataframe creation:
  - Extracted data is combined into a single dataframe containing job title, company, skills, location, salary, and experience.
  - This dataframe allows for further analysis and visualization of the collected data.

Web scraping allows for efficient data collection and analysis, enabling us to gain valuable insights into the data science job market.

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:

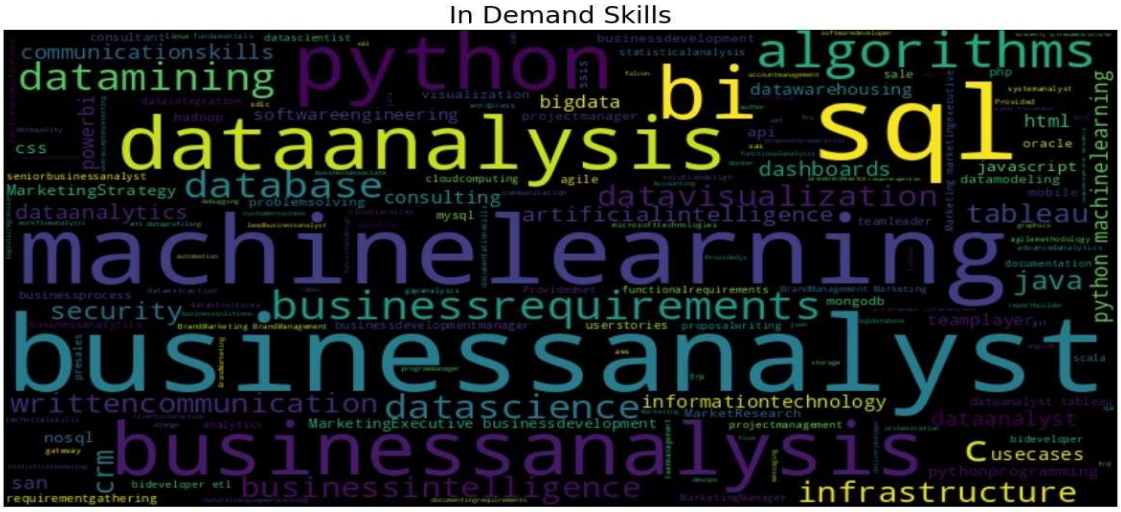
df.isnull().sum()

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.

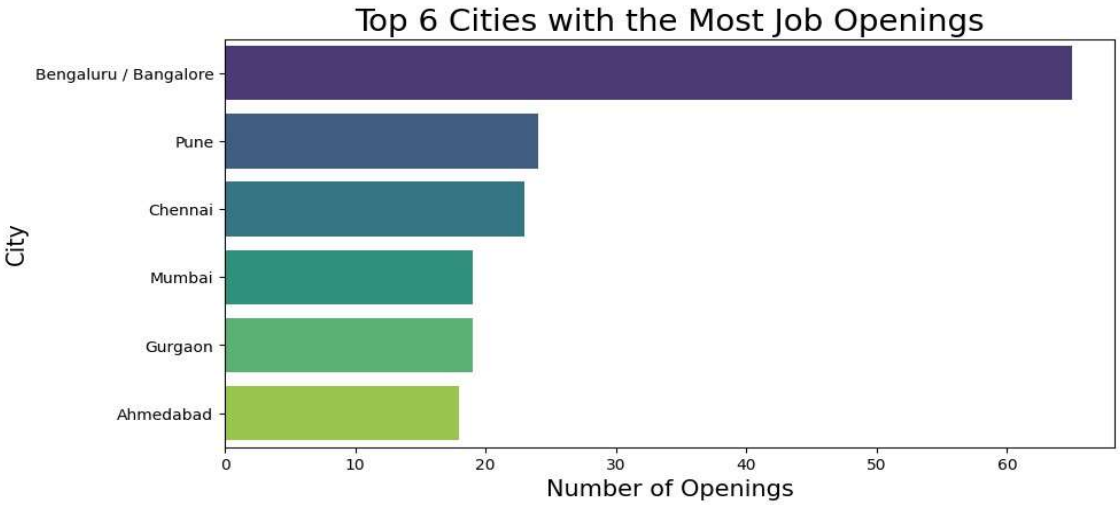
Visualizing In-Demand Skills with WordCloud:

WordClouds offer a swift and visually engaging overview of the most significant terms within a text. In this context, they efficiently summarize and visualize the skills currently sought after, as reflected in the 'Skills Required' column. The generated word cloud highlights the demand for skills such as Python, SQL, Machine Learning, Data Analysis, Data Mining, and Algorithms in the current job market.



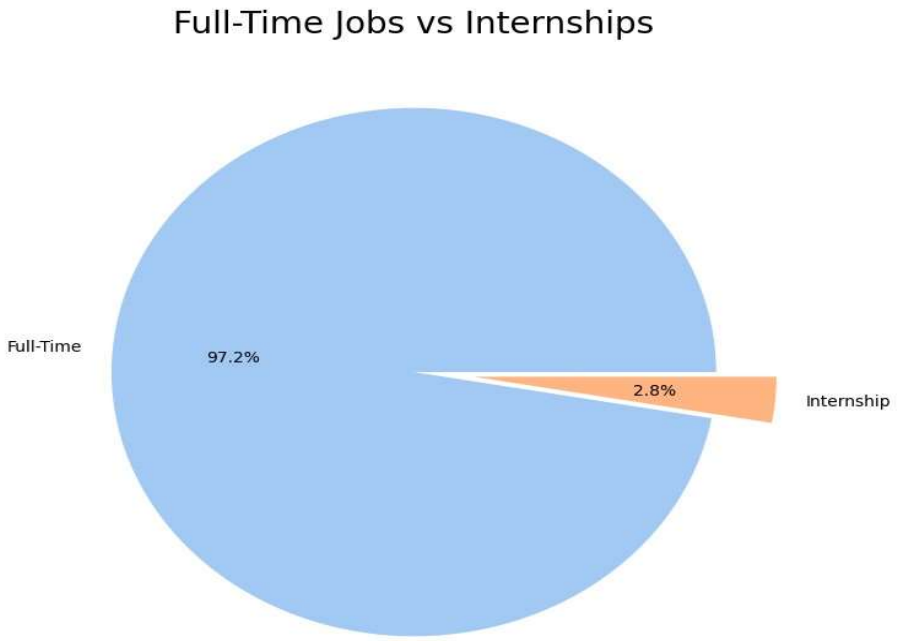
Top 6 Cities with the Most Job Openings in Data Science:

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.



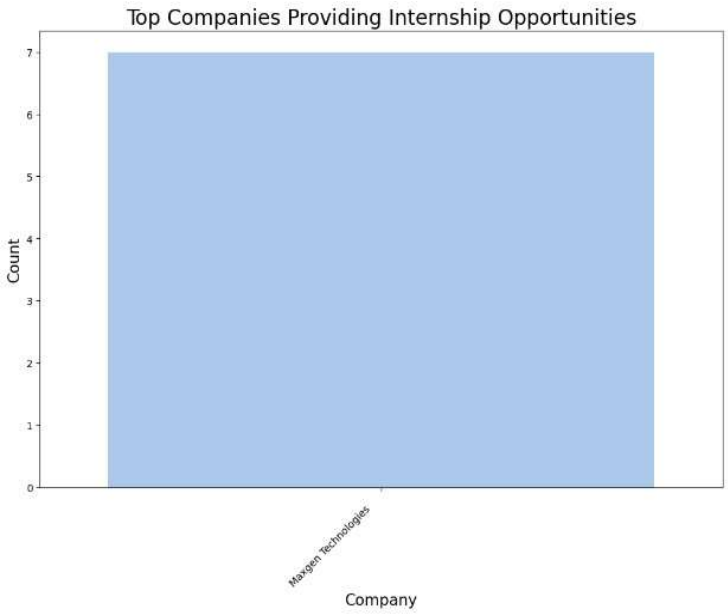
Comparison of Full-Time Jobs and Internships in the Job Market:

A pie chart serves as a visual representation of the percentage distribution within a dataset, effectively illustrating part-to-whole relationships. 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.



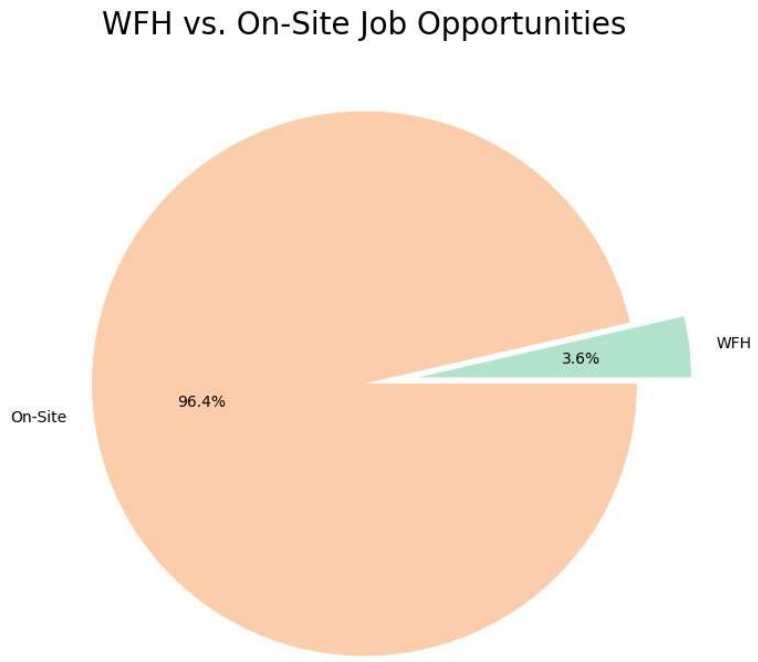
Top Companies Providing Internship Opportunities in the Job Market:

Bar charts effectively illustrate the frequency of occurrences across different levels of a categorical variable. In this instance, a bar chart is employed to identify the leading companies offering internship opportunities. The visualization highlights that Maxgen Technologies stands out as the top company providing internship opportunities in the current scenario.



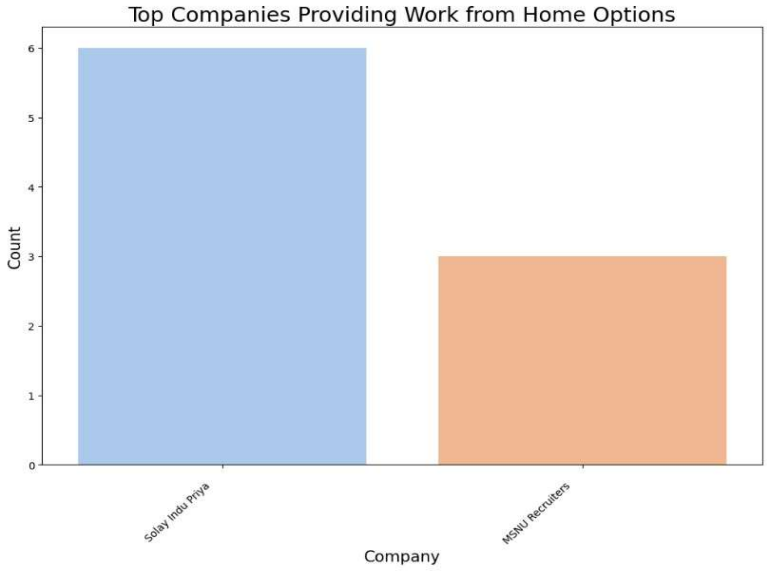
Comparison of Work from Home vs On Site Job Opportunities in the Job Market:

A pie chart serves as an effective tool to illustrate the percentage distribution within a dataset. In this context, 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.



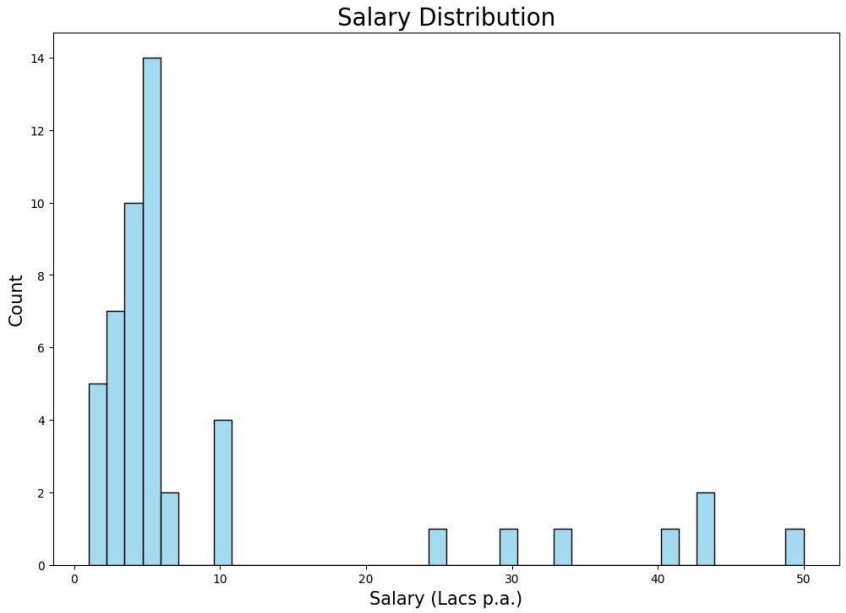
Top Companies Providing Work from Home Options in the Job Market:

Bar charts offer a visual depiction of frequency occurrences in various levels of a categorical variable. Therefore, a bar chart was employed to pinpoint the leading companies providing work-from-home opportunities at present, revealing that Solay Indu Priya and MSNU Recruiters are currently at the forefront.



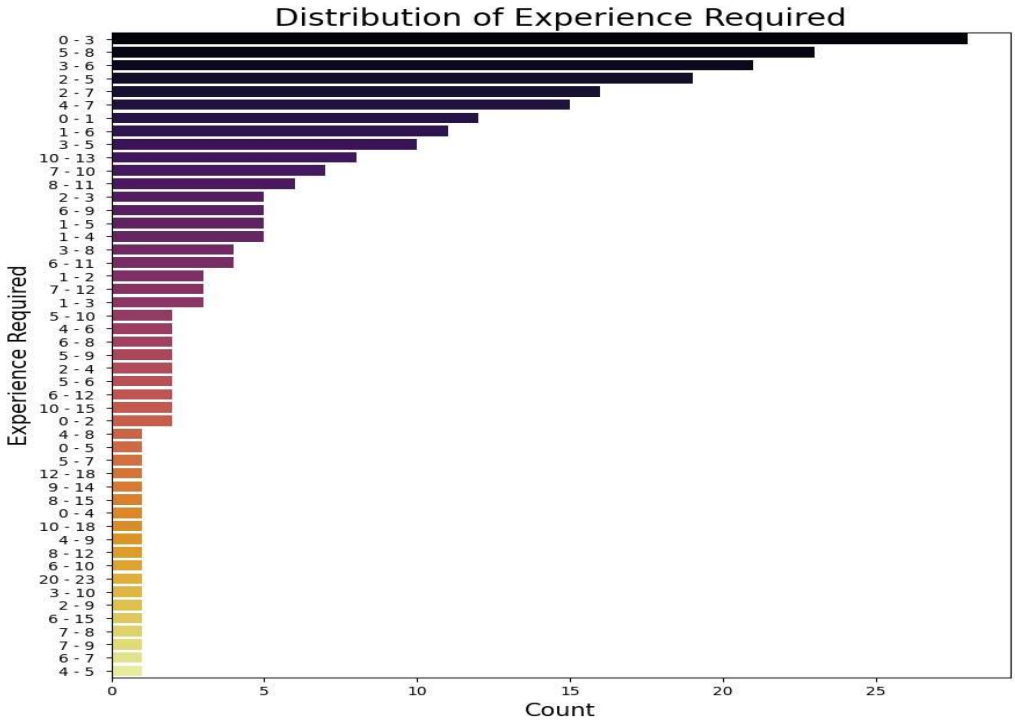
Salary Distribution in the Job Market:

Histogram chart enabled me to depict the current distribution of salary packages offered in the job market. Predominant clusters exist at 0-10 Lacs (entry-level), around 30 (mid-experience), and 40-50 (highly experienced).



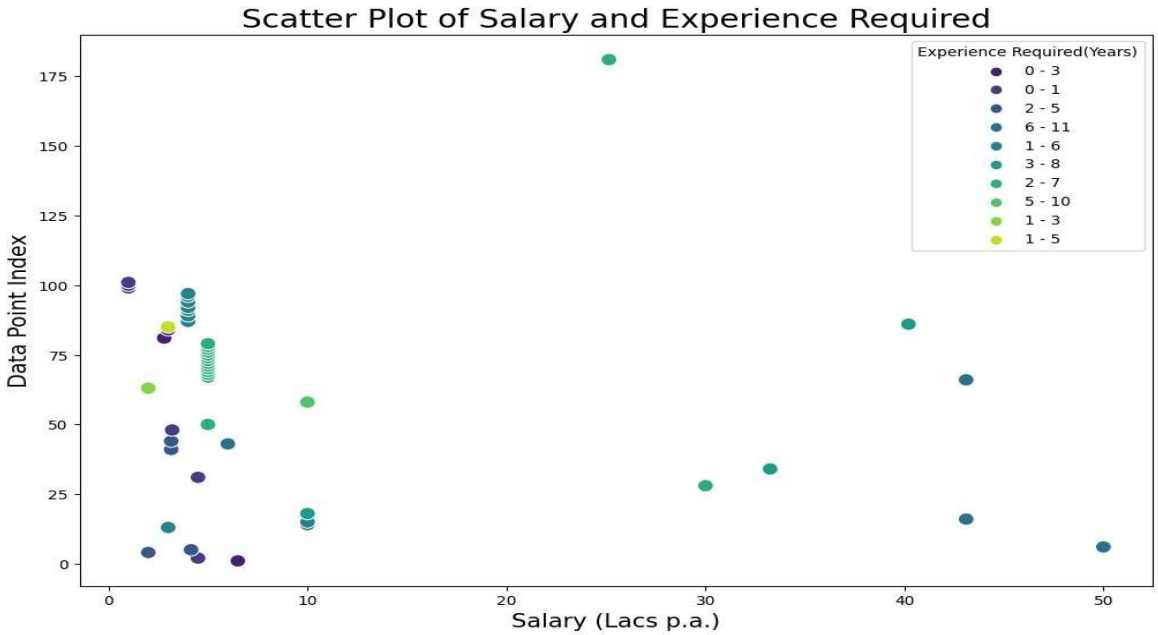
Experience Requirements in the Job Market:

Bar charts visually reveal the frequency of values across different levels of a categorical variable. In this instance, I utilized a bar chart to highlight the demand for varying experience levels in the job market, showcasing a preference for entry-level positions (0-3 and 0-1 years), mid-level roles (2-5, 36, 2-7, and 4-7 years), and experienced professionals (5-8 years).



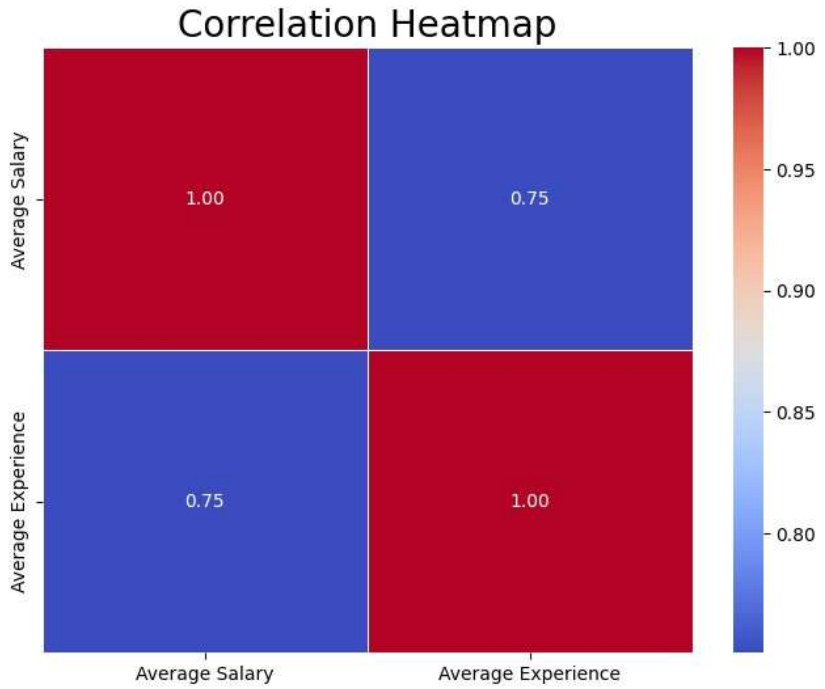
Analysing the Relationship Between Salary and Experience in the Job Market:

Scatter plots are an effective way to identify data patterns and correlations. By color-coding data points based on experience, this scatter plot quickly identifies salary trends relative to data point indices. The plot reveals distinct salary clusters, with prevalent entry-level positions (0-10 Lacs p.a.) and higher packages for experienced professionals (30-50 Lacs p.a.).



Correlation Heatmap

The correlation heatmap indicated a moderately strong positive relationship (correlation of 0.75) between Average Salary and Average Experience, signifying that, on average, as experience increases, so does the salary, and vice versa.



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.