Final Project Report

Trends and Skill Analysis in Data Science Job Postings

Submitted by: Siddhesh Wagh

Email: sidwagh9161@gmail.com

Master of Computer Applications (MCA)

[Institute of industrial and computer management and research, Nigdi, Pune]

AI & Data Science Internship Assignment

Tools Used: Python, Pandas, NumPy, Seaborn, Plotly, spaCy

July 2025

Project Objective

This project aimed to analyze real-world job postings for data science roles to extract actionable insights for students and job seekers. The key goals included:

- Cleaning and organizing job posting data.
- Extracting and categorizing in-demand skills using NLP.
- Identifying hiring trends based on job titles, companies, and locations.
- Presenting insights using visualizations and summaries.

Dataset Overview

Three datasets were used:

Dataset	Description
job_postings.csv	Contains metadata about job titles, company, location, job type
job_skills.csv	Skills extracted from job descriptions
job_summary.csv	Full textual job summaries

- **Total Records**: 12,217 job postings
- **Merged on**: job_link
- **Missing Values**: Cleaned, only minor missing entries in job_skills and job_location

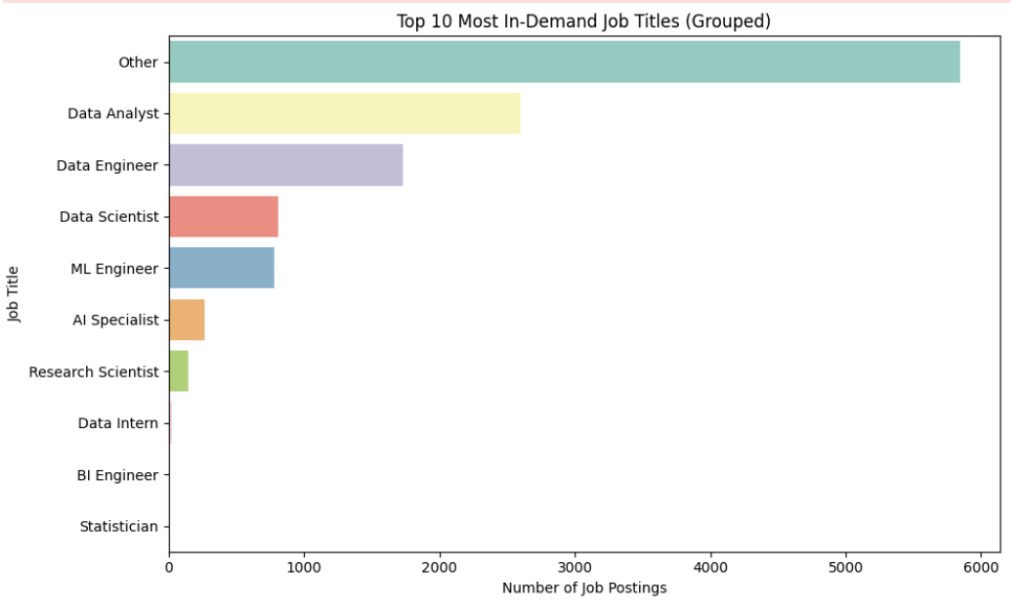
Data Cleaning

- Removed duplicates and handled null values
- Merged datasets using job_link
- Converted text to lowercase and stripped whitespace
- Standardized job levels into entry, mid, and senior

Exploratory Data Analysis (EDA)

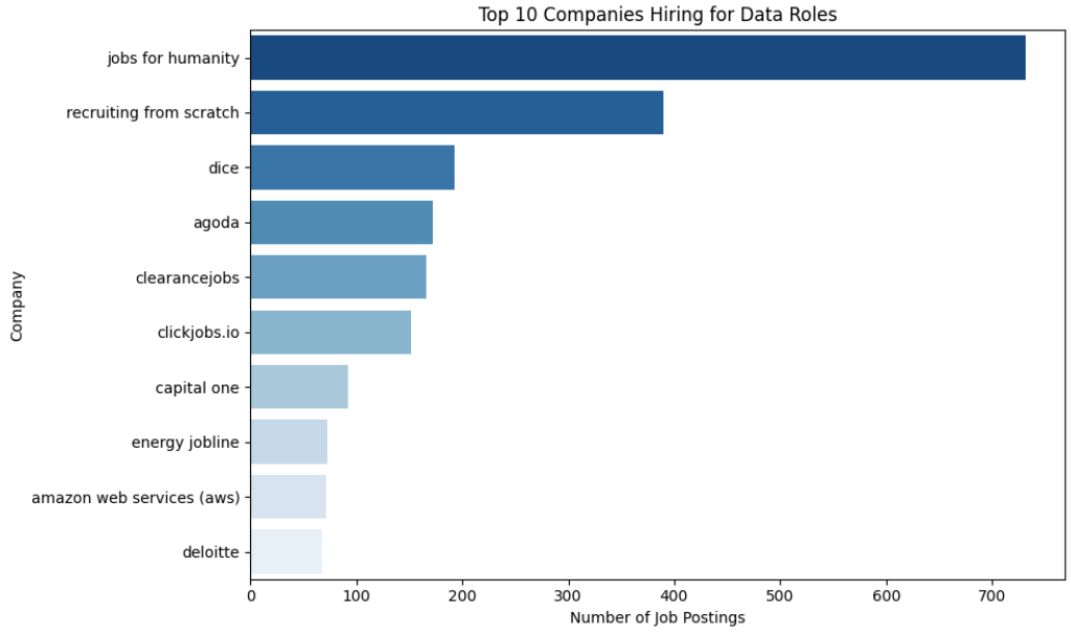
- **Top Job Titles**
  - Merged and grouped similar job titles (e.g., "Data Analyst", "Analyst - Data")

- **Top 10 Titles**: Data Scientist, Data Analyst, Machine Learning Engineer, etc.
- Barplot



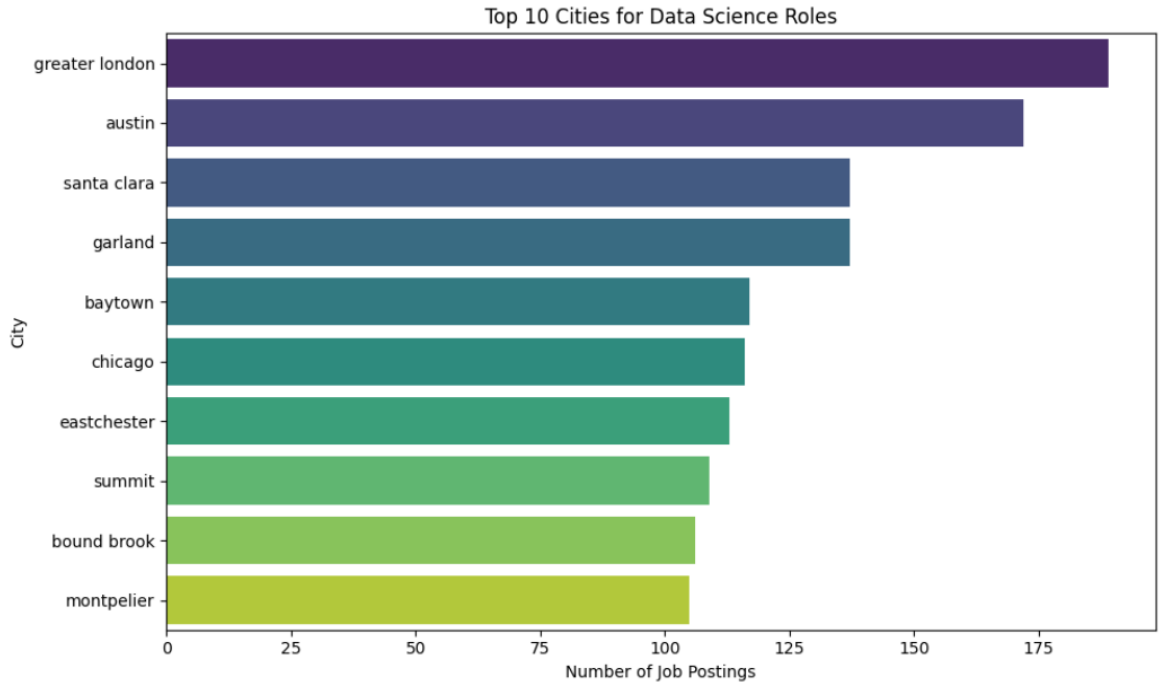
🔹 Top Hiring Companies

- Companies like **Amazon**, **Google**, **Accenture**, and **Meta** posted the most jobs.



🔹 Location Insights

- **Cities**: New York, San Francisco, Bengaluru were major job hubs.
- **Countries**: United States and India dominated the listings.
- Used Plotly for choropleth visualization of global trends.

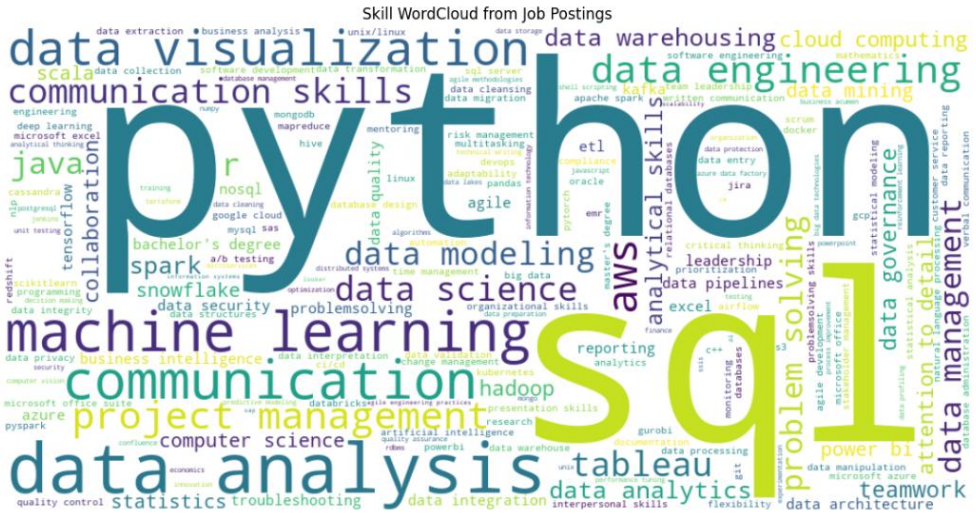


Skill Extraction & NLP

- Used spaCy and keyword tokenization on job_skills and job_summary
- Created frequency distribution of skills
- Separated **technical** vs **soft skills** (optional)

🔹 **WordCloud**

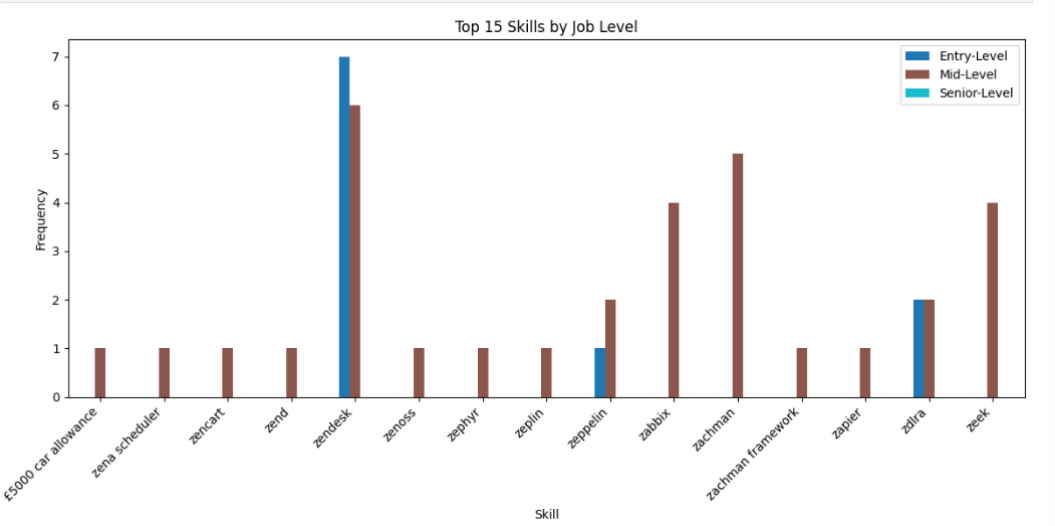
- Showed most frequent skills visually (e.g., Python, SQL, Machine Learning)



Skill Demand Across Job Levels

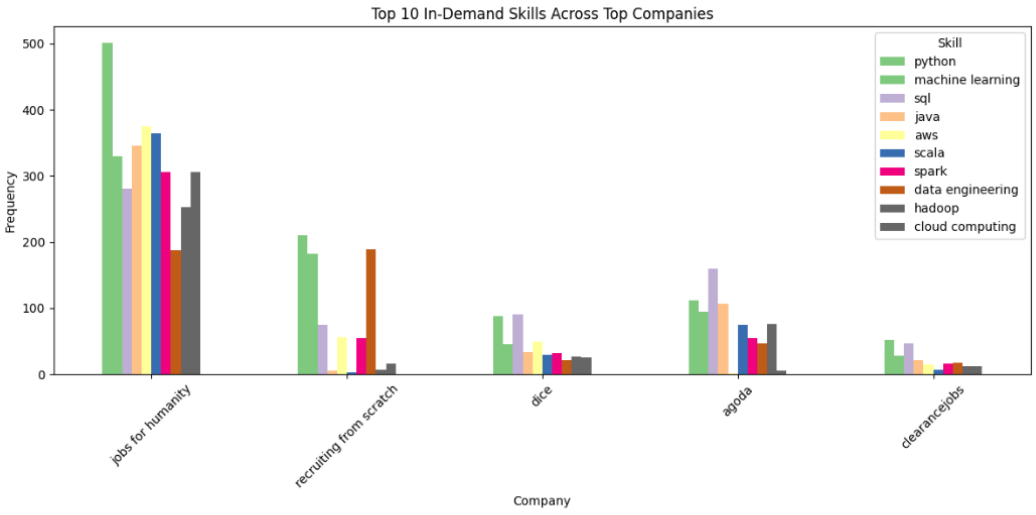
Job Level	In-Demand Skills
Entry	Excel, SQL, Python
Mid	Machine Learning, Tableau, Cloud
Senior	Deep Learning, Leadership, Spark

- Plotted a **comparison bar chart** across levels.



Skill Demand Across Companies

- Grouped job postings by company
- Identified which skills each company emphasizes
- Found specialization patterns (e.g., Meta focused on AI/ML, TCS on BI tools)



[ ]:

Key Insights

Category	Insight
Skills	Python, SQL, and Machine Learning are the top demanded skills
Job Titles	Data Scientist and Data Analyst dominate hiring
Location	US and India are top job markets
Company	FAANG companies lead in hiring volume
Job Level	Senior roles demand advanced tools and leadership

Tools Used

- **Python**: Pandas, NumPy, Matplotlib, Seaborn, Plotly
- **NLP**: spaCy
- **WordCloud**: WordCloud package
- **Environment**: Jupyter Notebook

Conclusion

This project has equipped the candidate with practical data science experience in:

- Cleaning and analyzing messy real-world datasets
- Applying NLP techniques for skill extraction
- Creating job market intelligence insights
- Communicating findings visually and analytically

It demonstrates a strong ability to handle workforce-related data in a business analytics setting.