GennieSphere
Magic. Innovation. Transformation

Assignment : AI & Data Science Intern

Title: Trends and Skill Analysis in Data Science Job Postings

Duration: 10 Days

Tools Allowed:

- Python (Required)
- Jupyter Notebook / Google Colab
- Pandas, NumPy
- Matplotlib / Seaborn / Plotly
- spaCy / NLTK (for NLP)
- WordCloud (optional)
- GitHub (for code hosting)

Objective:

Build a data-driven application that:

- Collects and cleans real-world job posting data for data science roles,
- Extracts and categorizes required skills using natural language processing (NLP),
- Identifies trends in job titles, skill requirements, hiring companies, and job locations, and
- Generates visual insights to help learners or job seekers understand current demands in the data science job market.

This project demonstrates how data science can be applied to real-world career intelligence, recruitment analytics, and skill-gap identification for professionals and students.

Task Details:

Day 1: Project Setup

- Download and explore the dataset
- Understand the structure and columns
- Check for missing values, data types, and quality
- Set up GitHub repo or Google Colab notebook

Output: Data overview summary, schema notes, project folder setup

Day 2: Data Cleaning

- Remove duplicates, handle nulls
- Standardize text fields (lowercase, strip whitespace)
- Filter for only data-related jobs (optional)
- Save clean CSV for reuse

Output: Cleaned dataset saved, notebook explaining cleaning steps

Day 3: Top Job Titles

- Analyze and visualize most in-demand job titles
- Group similar titles using NLP or string match ("Data Analyst", "Analyst - Data")

Output: Bar plot or pie chart of top 10 job titles

Day 4: Top Companies Hiring

- Find companies posting the most jobs
- Filter by region or seniority level if applicable

Output: Ranked table and visual of top hiring companies

Day 5: Location Insights

- Determine the most common job locations (city/country)
- Map using a geographic visualization (optional: plotly.express.choropleth)

Output: Map or chart showing top job hubs for data roles

Day 6: Skill Extraction (Text Mining)

- Tokenize and clean job description text
- Extract skills using keyword matching or NLP (use spaCy, or predefined skill list)
- Separate soft vs technical skills (bonus)

Output: WordCloud or bar chart of most common skills

Day 7: Skill Trends Across Job Levels

- Segment job postings by level (e.g., entry-level, senior)
- Compare required skills per level
- Use visual comparisons

Output: Comparative skill charts for each job level

Day 8: Skill Demand Across Companies/Industries

- Explore how skill requirements vary by company or industry
- Group and compare most requested skills across categories

Output: Multi-category comparison plot

Day 9: Final Insights & Report

- Summarize key trends:
    - Most in-demand skills
    - High-growth job titles
    - Regional hubs
- Write short, structured insights

Output: PDF report or Jupyter Markdown Summary

Day 10: Final Submission + GitHub

- Push code, visuals, and final report to GitHub
- Organize README with:

- Project overview
- Tools used
- Key findings
- How to reproduce

Output: GitHub link + ZIP folder (if required)

Deliverables:

- Cleaned dataset (CSV or notebook step)
- Jupyter/Colab notebook with:
  - Cleaning
  - Analysis
  - Visualizations
- A short report (Markdown or PDF)
- GitHub repo (optional but preferred)

Considerations:

- Cite the Kaggle dataset clearly
- Do not publish personal/private info (if any)
- Document code with comments and markdown cells
- Use consistent variable naming and plots

Submission Guidelines:

- Submit via GitHub (preferred) or ZIP folder
- Share path of data source
- Include report with visualizations
- Include final notebook with outputs
- If deployed (e.g., Streamlit), share public link
- Submit your assignment using this Google Form : **https://forms.gle/u2RFMrTCgZHDTATb6**

Outcome:

By the end of this project, freshers will gain practical experience in:

- Data cleaning, wrangling, and preprocessing of structured and unstructured job posting data,
- Exploratory data analysis (EDA) to uncover hiring trends and skill demands,
- Natural language processing techniques for skill extraction from job descriptions,
- Visualizing geographic, skill-based, and industry-based job trends, and
- Presenting insights through charts, word clouds, and dashboards.

This project will showcase their skills in data analysis, job market intelligence, and visualization — making them job-ready with practical exposure to analyzing workforce datasets in a business context.

Connect with us:
LinkedIn | Instagram | Twitter | Website| Youtube

📩 If you have any concerns, please reach out to us at **internship@genniesphere.ai**