InternSight: Internship Market Intelligence

By: Disha Tarlekar
Domain: Data Analytics / Web Scraping / Python

1. Project Overview

This project focuses on **extracting real-world internship data** from the Internshala website using **Python-based web scraping**.
After collecting raw data, it is **cleaned, transformed, and analyzed** using Pandas, followed by **visual insights** using Matplotlib/Seaborn.

The goal of this project is to understand **market trends**, including internship availability by location, stipend distribution, and industry patterns.

2. Objectives

- Scrape internship listings directly from the web
- Clean and preprocess the extracted dataset
- Convert unstructured text-based information into structured columns
- Analyze patterns in **location, stipend, duration, roles, companies**
- Create meaningful **visualizations**
- Generate insights useful for students or job seekers

3. Tools & Technologies Used

- **Python**
- **Libraries:** Requests, BeautifulSoup, Pandas, NumPy
- **Visualization:** Matplotlib, Seaborn
- **Data Storage:** CSV file
- **Jupyter Notebook**

4. Data Scraping Process

4.1 URL Collection

A Python script was written to automatically:

- Collect internship listing URLs from Internshala
- Filter unnecessary elements
- Store all URLs inside a list for structured scraping

4.2 Data Extraction

For each internship link, the script extracted:

- Job Title
- Company
- Internship Location
- Monthly Stipend
- Duration
- Full job description
- Apply link (URL)

All scraped data was stored in a structured Pandas DataFrame.

---

5. Data Cleaning Steps

Raw web data contains inconsistent formats.
To prepare it for analysis, the following steps were applied:

5.1 Cleaning Columns

- Removed empty rows
- Standardized title, company, and location text
- Converted stipend to numeric format
- Extracted numeric values from duration (e.g., "6 Months" → **6**)
- Removed duplicates (based on URL)

5.2 Final Dataset Export

A final cleaned dataset named:
internshala_clean_final.csv
was generated for analysis.

---

6. Exploratory Data Analysis (EDA)

Multiple visualizations were created to identify insights:

6.1 Internship Distribution by City

- Bar Chart

  To identify the top city offering the highest number of internships.

6.2 Stipend Analysis

- Distribution Plot
- Comparison across roles
  Gives an idea of stipend ranges across internships.

6.3 Duration Analysis

Shows which internship durations (1, 3, 6 months) are most common.

6.4 Role-wise Trends

Which fields offer the most internship opportunities.

---

7. Key Insights

- Major internships are concentrated in **Mumbai.**
- Most common durations are **3 months and 6 months.**
- Stipend varies widely, with median stipend values around moderate ranges.
- High internship demand is seen in **Data Science, Web Development, Business Analytics, and Marketing.**

---

8. Challenges Faced

- Handling dynamic website content
- Cleaning inconsistent text information
- Extracting correct numeric values from messy strings
- Avoiding duplicate URLs

---

9. Applications of the Project

- Helps students find internship trends
- Useful for market research
- Demonstrates end-to-end data pipeline creation
- Shows skills in **web scraping, cleaning, data analysis, and visualization**

---

10. Conclusion

This project successfully demonstrates the complete **data analytics lifecycle**, from collecting raw data to generating insights.
It highlights core skills needed for a **Data Analyst**, including:

- Web scraping
- Data cleaning
- Exploratory analysis
- Visualization
- Insight generation

The final dataset and charts can be used to build Power BI/Tableau dashboards as well.
