Project Title:

End-to-End Web Scraping and Data Analysis on Internshala Internship Listings

Tools Used:

Python,
Selenium, BeautifulSoup,
Pandas, NumPy, Matplotlib, Seaborn

EXECUTIVE SUMMARY

This project demonstrates the complete lifecycle of a real-world data project starting from data extraction, cleaning, processing, visualization, and finally insight generation.

Data was collected directly from Internshala, focusing on Data Analyst Internships. A custom scraper was built using Selenium to navigate dynamic elements and BeautifulSoup for HTML extraction. A total of 50–80 internship listings were extracted.

The cleaned dataset was used to analyze patterns in location, companies, duration, and stipends. This project highlights key skills required in industry-level data roles.

OBJECTIVES

- Automate data extraction from a dynamic website.
- Clean and preprocess raw scraped data.
- Convert unstructured text into structured variables.
- Analyze internship trends for Data Analyst roles.
- Build visualizations for insights.
- Produce an industry-ready project report.

Key Questions Answered

- Which city have the most Data Analyst internship?
- What stipend range is most common?
- Which companies post the highest number of roles?
- Does stipend depend on internship duration?

Technologies Used

- Python
- Selenium WebDriver
- BeautifulSoup
- Pandas & NumPy
- Matplotlib & Seaborn

Scraping Methodology

- Selenium loads dynamic pages.
- BeautifulSoup parses content.
- Extracted fields:
  - Internship Title
  - Company
  - Location
  - Stipend
  - Duration
  - Internship URL

DATA CLEANING and FEATURE ENGINEERING

Cleaning Steps Applied:

- Removed duplicates based on internship URL
- Standardized location names
- Parsed stipend into numeric monthly amount
- Extracted duration using regex
- Cleaned title and company names
- Handled missing values
- Created 3 new derived columns:
  - Location_Clean
  - Stipend_Monthly
  - Duration_Clean

Challenges Solved

- Dynamic website delays
- Inconsistent stipend formats
- Irregular multi-location listings
- Mixed duration formats (weeks/months)

VISUALIZATION and EDA

Charts Created:

1. Top City Offering Internship
2. Stipend Distribution
3. Top Companies by Internship Count
4. Duration Distribution
5. Stipend vs Duration Scatterplot
6. Correlation Heatmap

Key Patterns Found:

- Mumbai, Bangalore, and Delhi show the highest opportunities.
- Most stipends fall between ₹5,000–₹15,000.
- Stipend and duration show very low correlation.

INSIGHTS and BUSINESS ANALYSIS

1. City-Based Opportunities
Internships are highly concentrated in Tier-1 cities:

- Mumbai

2. Companies Posting Frequently
Certain tech startups and analytics firms post the majority of roles, indicating high hiring demand.

3. Stipend Trends

- 60–70% internships offer ₹5k–15k.
- Very few roles exceed ₹25k.
- Some unpaid listings still exist.

4. Duration Trends
Majority range from:

- 1 month
- 2 months
- 3 months

5. Correlation Understanding
Stipend does not depend on duration
→ Companies prioritize skill fit over duration.

CONCLUSION

This project covers:

- Real-world scraping
- Data engineering
- EDA
- Analytical storytelling

It proves strong understanding of:

- ✓ Python automation
- ✓ Web scraping
- ✓ Data preprocessing
- ✓ Visualization
- ✓ Insight communication

FUTURE SCOPE

- Deploy scraper weekly for live trends
- Extend to multiple job platforms
- Build a dashboard using Power BI/Tableau
- Add ML model to predict stipend based on features

WHAT I LEARNED

- Handling dynamic websites using Selenium
- Parsing HTML with BeautifulSoup
- Writing robust scraping scripts
- Cleaning messy real-world data
- Visualization best practices
- Analytical phrasing for reports
- Building end-to-end data projects
- Preparing job-ready data analytics documentation

Thank You