

Project Proposal

The project is titled Book Price Web Scraping, Cleaning, Analysis, and Visualization Pipeline. The team consists of one member, Qiyan Chen (qiyanche@usc.edu, GitHub username: qiyanche, USC ID: 1301370183).

1. What problem are you trying to solve?

The goal of this project is to build an end-to-end data pipeline that automatically collects, cleans, analyzes, and visualizes book price data from an online source.

Book prices vary widely across categories, ratings, and individual products, and manually collecting price information is time-consuming and error-prone.

This project aims to solve the following problems:

- How to automatically gather structured price information from a website.
- How to clean inconsistent text/HTML data into usable structured format.
- How to summarize pricing patterns using descriptive statistics.
- How to visually communicate pricing distributions and identify high-priced items.

The project demonstrates practical skills in web scraping, data engineering, and exploratory data analysis.

2. How will you collect data and from where?

Data will be collected directly from <http://books.toscrape.com>, a publicly available website designed for web scraping practice.

The collection method:

- Use Python requests to send HTTP GET requests to book listing pages.
- Use BeautifulSoup to parse HTML content and extract:
book title, price, rating, product page URL, snapshot timestamp
- Save raw extracted data as timestamped JSON files in /data/raw/

This ensures that the dataset is fully collected through Python code, not from pre-downloaded files, meeting course requirements.

3. What analysis will you do and what visualizations will you create?

Data Analysis

Using Pandas and NumPy, the project will compute:

- Global descriptive statistics (mean, median, std, min, max, percentiles)
- Price distribution across all books
- Per-product aggregated statistics:
 - a. count of observations
 - b. minimum price
 - c. maximum price
 - d. average price

Outputs will be stored as:

- summary_stats.json
- metrics_by_product.csv

Data Visualization

Using Matplotlib, the following visualizations will be created:

1. Histogram of book prices — shows distribution and skewness
2. Boxplot of prices — highlights variation and outliers
3. Top 10 most expensive books bar chart — provides comparison across items

All plots will be saved into the /results/ directory.

4. Expected Outcomes

By completing this project, I will deliver:

- A working Python web scraper
- Cleaned and structured dataset
- Statistical summary of book pricing patterns
- A set of clear and meaningful visualizations
- A replicable data pipeline demonstrating web scraping, cleaning, analysis, and visualization

This project showcases practical data engineering and analysis skills and results in a reusable framework that can be adapted to other price-tracking applications.