Data Science Tools project

Team: Omar Mohamed Fahmy Mahmoud _ 23011389

Amr Khalid Atia _ 23011395

Fady Anter _ 23011404

Abdollah Ibrahim Mohamed _ 23011351

http://books.toscrape.com**Website used:**

Target Data **to extract and used for visualization:** Book titles, prices, availability, category, ratings, and product details.

Libraries used to do the required tasks:

requests, BeautifulSoup (bs4), pandas, re, time, matplotlib, seaborn, pyodbc, streamlit.

Project Idea & Goal:

The goal is to extract and analyze book data from the online bookstore “Books to Scrape”. The analysis will provide insights into price distribution, book availability, popular categories, and rating trends. It will also involve cleaning and organizing textual data using regular expressions and Visualizations by plots and diagrams to show the relationships between the features.

Project Plan:

1.Data Extraction

#Scrape book information across all categories and pages using requests and BeautifulSoup.

#Save raw data into a structured format like CSV or JSON.

2.Data Cleaning & Processing

#Remove invalid or duplicate entries.

#Standardize data formats and extract key text details (e.g., prices and ratings) using Regex.

3.Data Analysis

#Compute descriptive statistics (average price per category, rating distributions, etc.).

#Explore patterns such as category-wise trends or availability insights.

4.Data Visualization

#Use libraries like matplotlib and seaborn to generate:

- Boxplot of book counts per category.
- Countplot of rating distributions.
- Histplot showing price variation.

-Scatterplot showing price vs availability.

5.Data Storage

#Save the cleaned data to a MySQL database for easy retrieval and further exploration.

6.Bonus Task – Streamlit Web App

#Build an interactive dashboard using Streamlit.

#Users will be able to view analysis results, filter by category or rating, and explore visualizations dynamically.

#This enhances the presentation and accessibility of the findings.