**COS 598: Introduction to Data Science**

**Fall 2021**

**Homework Assignment 1**

Assigned: September 9, 2021

Due: September 23, 2021

*Submission instructions*:

You must submit your Python code in a .py file, and include a brief write-up about what you did in your code to complete the tasks in this assignment, and what results you obtained after running your code. You must also submit the .csv files from both tasks.

The file(s) must be uploaded to the Brightspace submission site for Homework 1.

Total: 100 points.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Task 1. Web Scraping (60 points).

In this task, you will scrape the book depository website’s bestsellers pages:

<https://www.bookdepository.com/bestsellers>

The above link will take you to the *first page* of the bestsellers on the book depository website. However, there are *more than one page* of bestsellers on the website. Your task is to scrape *all* pages of the bestsellers on the website, and collect the following information about *each* bestselling book: title, author, publication date, format, *current* price, and the *original* price. You should construct a Pandas’ dataframe to hold the above information for each bestselling book. Your dataframe should look like:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | title | author | publication\_date | format | current\_price | original\_price |
| 0 | It Ends With Us: The most heartbreaking novel you'll ever read | Colleen Hoover | 02 Aug 2016 | Paperback | US$16.79 | US$16.79 |
| 1 | The Beekeeper of Aleppo | Christy Lefteri | 01 Apr 2020 | Paperback | US$10.77 | US$12.46 |
| … | | | | | | |

Note: There are missing values (e.g., some books do not have a price listed).

Save the dataframe as a .csv file named bestsellers.csv.

Task 2. Data Wrangling and Exploration Using Pandas (40 points).

1. Process the above dataframe such that the publication date is split into three integer columns: publication year, publication month, and publication day. Also, remove “US$” from the current price and the original price column. In addition, you should remove any row that has a missing value in any column. Your dataframe should eventually look like:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | title | author | publication\_year | format | current\_price | original\_price |
| 0 | It Ends With Us: The most heartbreaking novel you'll ever read | Colleen Hoover | 2016 | Paperback | 16.79 | 16.79 |
| 1 | The Beekeeper of Aleppo | Christy Lefteri | 2020 | Paperback | 10.77 | 12.46 |
| … | | | | | | |

1. Save your dataframe as a .csv file named bestsellers-cleaned.csv.

Use the cleaned dataset to do the following:

1. Find the titles, authors, and current prices of all bestselling books published in 2019 or later. Include your answer in your write-up.
2. Create a scatterplot of the current price versus the original price for the bestselling books. You must include your scatterplot in your write-up.
3. Create a histogram to visualize the distribution of the current prices of the bestselling books. You must include your histogram in your write-up.
4. What is the average current price of the bestselling books for *each* publication year? Include your answer in your write-up. Create a bar plot to visualize the average current price of the bestselling books for *each* publication year, and include the bar plot in your write-up.