CW1 – Amazon Bestsellers Analysis with PySpark

- Due 19 Mar by 12:00
- Points 30
- Submitting an external tool
- Available 17 Feb at 0:00 - 26 Mar at 12:00

Weighting %:	30	**Submission deadline (for students):**	See Canvas
Authorship:	Individual	**Target date for returning marked coursework:**	
Tutor setting the work:	Ashley Spindler	**Number of hours you are expected to work on this assignment:**	

This Assignment assesses the following module Learning Outcomes (from Definitive Module Document):

1. be able to write data processing pipelines that exploit Apache Spark

2. have knowledge of and understand how Apache Spark can be used to handle and process data

Assignment Tasks:

In this assignment you will be tasked with exploring a dataset containing the Top 50 best-selling books from Amazon between 2009-2019. You should complete the exercises presented in the Google Colab Notebook below. This assignment will be graded using CodeGrade.

Notebook: PySpark_Assignment_Feb24.ipynb (https://herts.instructure.com/courses/110656/files/8360850?wrap=1) ⤓ (https://herts.instructure.com/courses/110656/files/8360850/download?download_frd=1)

Dataset: AmazonBooks-1.csv (https://herts.instructure.com/courses/110656/files/8360851?wrap=1) ⤓ (https://herts.instructure.com/courses/110656/files/8360851/download?download_frd=1)

Exercise 1 (5 Marks): Find the authors with the most entries in the bestseller's lists, find the number of unique titles for each, the average rating, total number of reviews, and highest position in the ranking.

Exercise 2 (5 Marks): For fiction and non-fiction books, find the average and total number of reviews for the top 10, 25, and 50 of the bestsellers lists, in each year.

Exercise 3 (10 Marks): For each year, find the average price of a fiction and non-fiction book in the top 10, 25 and 50 of the bestsellers lists.

Exercise 4 (10 Marks): For free books—where the price is zero—fine the number of unique titles and authors. Compare the average rating and number of reviews in each year between free and priced books.

Submission Requirements: Submission for this assignment is done via CodeGrade. You should submit your notebook via the upload option. You may submit your work an unlimited number of times.

Marks awarded for:

Type of Feedback to be given for this assignment:

Additional information:

· Regulations governing assessment offences including Plagiarism and Collusion are available from https://www.herts.ac.uk/__data/assets/pdf_file/0007/237625/AS14-Apx3-Academic-Misconduct.pdf (https://www.herts.ac.uk/__data/assets/pdf_file/0007/237625/AS14-Apx3-Academic-Misconduct.pdf) (UPR AS14).

· Guidance on avoiding plagiarism can be found here: https://herts.instructure.com/courses/61421 (https://herts.instructure.com/courses/61421) (see the **Referencing** section)

· For **postgraduate modules**:

o a score of 50% or above represents a pass mark.

o late submission of any item of coursework for each day or part thereof (or for hard copy submission only, working day or part thereof) for up to five days after the published deadline, coursework relating to modules at Level 7 submitted late (including deferred coursework, but with the exception of referred coursework), will have the numeric grade reduced by 10 grade points until or unless the numeric grade reaches or is 50. Where the numeric grade awarded for the assessment is less than 50, no lateness penalty will be applied.

This tool needs to be loaded in a new browser window

Load CW1 – Amazon Bestsellers Analysis with PySpark in a new window