BIA 6301 APPLIED DATA MINING

HOMEWORK ASSIGNMENT #4

General Instruction: The homework assignment is due one week after it is assigned (@ 5:45 PM on Tuesday). Assignment turned in after the due date and time will lose 2 points for every day late. No assignment will be accepted one week after it is assigned. Here’s a breakdown of the point distribution.

|  |  |  |
| --- | --- | --- |
| **Task** | **Points Possible** | **Submission Instruction** |
| GitHub repository | data & script = 8 points | Submit the hyperlink to the GitHub account in Blackboard for grading. |
| Markdown file with written commentary. | 3 grading parameters x 4 points = 12 points  The markdown file will be graded using the Written Assignment Grading Rubric. |
| ***Total*** | ***20 points*** |  |

**Gaining Familiarity with R Project, Git, and GitHub**

Work through the “A Very Brief Tutorial on R Project and Git without Using Command Line” posted under Module\_4/Class/docs in Blackboard.

**Case Study: Book displays at Flourish and Blotts**

The purpose of this assignment is to show how association rules can simulate the book recommendation feature on Amazon.com.

The manager of Flourish and Blotts bookstore is trying to decide the best ways to arrange the book displays for the upcoming summer season in order to increase sales. He gives you a large data set with 90,000+ historical sales transactions and asks you to provide the answers for three questions:

1. What are the best-selling titles?
2. If the manager has to create a book display to appeal to readers who belong to book clubs, what books should be included? He said that the “typical” book club audience would be someone who is reading titles featured by Oprah’s Book Club (<https://static.oprah.com/images/o2/201608/201608-obc-complete-list-01a.pdf>).
3. Can you recommend other books that he should include in display cases? The manager is adamant that your recommendations **do not** include the following:
   1. Books in a series (i.e. Girl with the Dragon Tattoo series would be an example). The manager already knows series books should be displayed together.
   2. The title “Wild Animus.” You were really surprised by this request and pressed the manager for an explanation. He replied that you should read this blog entry: <https://litreactor.com/columns/what-the-hell-is-wild-animus>

Instructions on How to Proceed:

Download **bookdata.tsv.gz** and **Bookbaskets.R** from Blackboard. The dataset is in tab separated format (.tsv). The .gz extension means the file is compressed due to its large size. If you want to view the raw data, you can download the pdf file titled “bookdata\_raw” from Blackboard. The Bookbaskets.R file has codes to import the bookdata.tsv.gz file and convert it into a sparse matrix. Use the .R file to get you started for this week’s homework assignment.

Create an R markdown file to address the manager’s questions. You are welcomed to embed R code chunks, visualizations and graphics, hyperlinks, and R outputs in the markdown file. Please provide detailed commentary in the markdown file. In particular, your commentaries need to address the following points:

1. What is the problem/question? How are you going to solve/answer it?
2. What analysis did you carry out to solve/answer the problem/question?
3. What are your recommendations to solve/answer the problem/question?

The standard you need to reach for this week is the following: If you look at the R markdown file six months from now, can you understand what you did and can you talk someone else through what you did? The learning goal for this week is to tell a compelling story with all the technical tools and written communication skills you have learned in previous week. Remember that you are casting a wider net now—talking to an audience that may include managers and technical people. So while jargons are acceptable, keep them to a minimum and explain/define your jargons.

I will grade the detailed R markdown file using the same Written Assignment Grading Rubric as previous weeks.

**Instead of the memorandum write up, you will be creating a detailed R markdown file this week.** I will grade the markdown file using the same Written Assignment Grading Rubric. Hence, you should make sure that you address all three grading parameters: a) state the problem/question & method(s) you used to solve it; b) discussion of what you did to solve the problem/question; and c) recommendations. Incorporate R codes, R outputs, and visualizations as needed to craft your R markdown file. The learning goal for this week is to tell a compelling story.

Here are some examples that may help you with the assignment:

<https://blog.nycdatascience.com/student-works/machine-learning/movie-rating-prediction/>

<https://www.kaggle.com/apryor6/d/apryor6/us-opiate-prescriptions/detecting-frequent-opioid-prescription>