School of Computer Sciences

CPC351/CPM351 Principles of Data Analytics

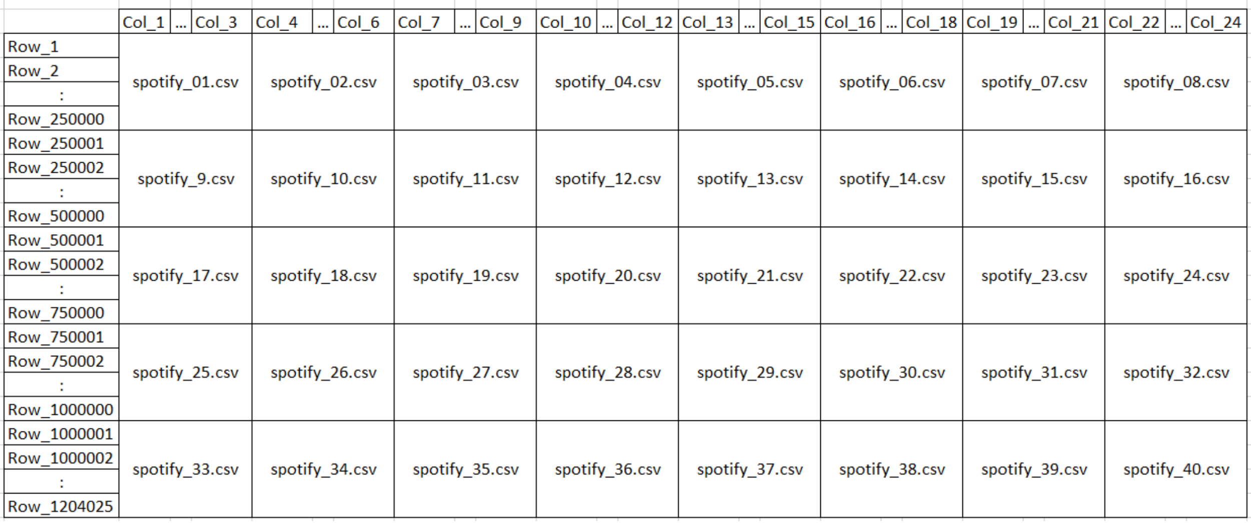
Academic Session: Semester 1, 2024/2025

Assignment 01 – Basic of R Programming

I. Questions

You are required to answer the following questions:

1. Download **tracks_features.csv** from eLearn@USM. It contains audio features for over 1.2 million songs. This dataset is taken from Kaggle website[1]. Reference for these audio features can be found in the Spotify website[2]. Since this is a dataset with about 1.2 million rows and 24 columns. You are required to write a R program to split the CSV file into 40 CSV files with the following settings:



	Col_1 … Col_3	Col_4 … Col_6	Col_7 … Col_9	Col_10 … Col_12	Col_13 … Col_15	Col_16 … Col_18	Col_19 … Col_21	Col_22 … Col_24
Row_1 Row_2 : Row_250000	spotify_01.csv	spotify_02.csv	spotify_03.csv	spotify_04.csv	spotify_05.csv	spotify_06.csv	spotify_07.csv	spotify_08.csv
Row_250001 Row_250002 : Row_500000	spotify_9.csv	spotify_10.csv	spotify_11.csv	spotify_12.csv	spotify_13.csv	spotify_14.csv	spotify_15.csv	spotify_16.csv
Row_500001 Row_500002 : Row_750000	spotify_17.csv	spotify_18.csv	spotify_19.csv	spotify_20.csv	spotify_21.csv	spotify_22.csv	spotify_23.csv	spotify_24.csv
Row_750001 Row_750002 : Row_1000000	spotify_25.csv	spotify_26.csv	spotify_27.csv	spotify_28.csv	spotify_29.csv	spotify_30.csv	spotify_31.csv	spotify_32.csv
Row_1000001 Row_1000002 : Row_1204025	spotify_33.csv	spotify_34.csv	spotify_35.csv	spotify_36.csv	spotify_37.csv	spotify_38.csv	spotify_39.csv	spotify_40.csv

You are required to use data frames in your implementation. After tracks_features.csv is split into 40 CSV files, you need to import each of these CSV files into R and combine them as a data frame named complete.

[1] https://www.kaggle.com/rodolfofigueroa/spotify-12m-songs

[2] https://developer.spotify.com/documentation/web-api/reference/#/endpoint-get-audio-features

2. Amazon and Walmart Case Study. Read and watch the video. Answer the questions given.

   Set in 2021, this case describes how Amazon and Walmart have been two of the most successful retailers in history and are responsible for changing the rules of the game in the retail industry in the US. Over the years, the two firms have perfected contrasting business models to enable their dominance in offline and online retailing. Walmart's model of low prices and strategic partnerships with suppliers redefined supply chain practices and lowered system costs through the adoption of information technology. Amazon's online model of convenience of shopping from anywhere, anytime, comprised a high-quality, user-friendly platform with a large product catalog and a widespread and reliable fulfilment infrastructure to deliver the orders quickly to the shopper. In recent years, the growing customer preference for omni-channel retailing, an integrated experience that seamlessly comprises digital and physical retail, has compelled the two companies to make substantive investments in developing capabilities and acquiring resources in what was hitherto the other's domain.

   Link to Video:
   https://www.youtube.com/watch?v=bnMjaiBZ3Jo&list=PLJqiHnBNjacrnUmg6tKXQ9tRGuuyM2tn-&index=40&ab_channel=5MinutesLearning

   i. The upheavals of 2020-2021, driven by the COVID-19 pandemic, fueled a surge in online shopping, significantly removing the gap between Amazon and Walmart. Using the values provided in the video, identify variables X and Y, and plot graphs comparing the revenue gains of the two companies. Provide your own analysis based on the plot.

   ii. In 2021, after COVID-19 struck, online grocery shopping significantly impacted Amazon's business landscape. They adapted to the changes, but over time, performance fluctuated based on certain items. Several variables made it challenging for them to maintain profitability in online grocery shopping. Online grocery retailers face difficulties related to **margins**, **customer price sensitivity**, and **delivery preferences**.

Write an R conditional statement that evaluates a store's profitability based on these three factors. Assign each factor a score and define the formula for Total score. Create a profitability classification:
(Note: You should assign each factor a score. You may make your own assumptions)

- High Profitability: Total score > 25
- Moderate Profitability: Total score 15–25
- Low Profitability: Total score < 15

iii. Given that Walmart operates *X* stores and clubs worldwide and serving *Y* million customers per week, write a conditional statement in R to categorize the average number of customers per store per week. Find out the values of *X* and *Y* from the video.

Define 3 different types of categories to classify the number of weekly customers per week. Calculate the average number of weekly customers per store, then write a conditional statement in R to assign the appropriate category. Print the category along with a message indicating the number of customers per week.

(Note: Provide your own assumptions when defining the 3 categories of customers per week).

iv. In 2005, Amazon introduced Amazon Prime, a game-changer in online shopping. The service led to rapid growth, expanding from *X* million members in 2011 to *Y* million in 2021. Find the values of *X* and *Y* from the video. Write an R function to calculate the number of members for any given year based on a linear growth rate. Use linear interpolation to estimate the number of members for a given target year. Test the function by using 2018 as the target year and verify the output. *(Note: Assume the membership numbers increase at a constant rate over time).*

3. Download the datasets given in the folder **Amazon Products.zip** from eLearn@USM. It contains 108 csv files of different categories of Amazon products. Each csv file consists of 9 columns and each row has product details accordingly. This dataset is taken from Kaggle website[3].

   i. Import all datasets into RStudio and combine them into a single csv file called Amazon Products All. How many samples are in the datasets?

   ii. Extract the manufacturer information from the 'name' column and insert a 'manufacturer' column immediately after the 'name' column.

   iii. The columns **actual_price, discount_price, no_of_ratings** and **ratings** have incorrect data types. They are currently stored as objects, but they should be of type int or float. Correct the data types accordingly. Explain the steps you took and provide the R code.

   iv. Create a new column that contains the discount percentage. Write a relevant formula to calculate the discount percentage and implement it in R.

   v. Group the data by the 'manufacturer' column and summarize the manufacturer with highest percentage discount of the item.

   vi. Write a function to categorize manufacturers into different rating values. First, filter the ratings to include only values in the range of 0.0 to 5.0. Clean the 'ratings' column as needed. Use the following categories, and assign your own variable names:

      - High: Ratings greater than 3.0
      - Medium: Ratings between 1.5 and 3.0
      - Low: Ratings below 1.0

   vii. Identify the products with the highest and lowest prices. Which manufacturer offers the product with the highest price? Analyze the relationship between the discount (calculated in iv) and the number of ratings. Explain your findings.

   viii. Determine the product with the highest sales. Sort the data in descending order of sales. What conclusions can you draw from your findings?

[3] https://www.kaggle.com/datasets/lokeshparab/amazon-products-dataset/data

II. Submission

This is a group assignment (a group of four members). The member grouping is done via eLearn@USM.

You are required to submit a zip/rar package which consists of the following items to the eLearn@USM:

- R script (in .R format).
- An assignment report not more than 12 pages (in pdf format). Only the sample output screen shots and relevant explanation/write-up/description are expected. Also, a cover page which contains your details must be included in your assignment report.

The zip/rar package must be named according to the following notation: CPC351_CPM351_[GroupNumber]_A01. For example, for Group03, they must name the zip/rar package as CPC351_CPM351_Group03_ A01.

One of the group members is required to submit the zip/rar package. Kindly communicate with your group member before the submission to avoid any miscommunication.

The submission deadline 24 November 2024 (Sunday), 23:59 p.m. Failure to submit the assignment will be a disadvantage to you.

III. Grading Rubric

This assignment will be graded according to the grading rubric as shown in Table 1. The total will be scaled to 7% of your overall grade.

Table 1: Assignment 01 grading rubric.

	Good (3)	Satisfactory (2)	Poor (1)	Fail (0)
Question 1 (20%)	• Meet all the requirements. • The R programme can be executed, and correct outputs are shown. • Clear and detailed comments are added to scripts with excellent clarity. • The report includes the screen shots, and explains the results with excellent clarity, comprehensiveness, and organization. • Discussions are well focused and important points are included.	• Partially meet the requirements. • The R programme can be executed, and partially correct outputs are shown. • Adequate comments are added to scripts with satisfactory clarity. • The report includes the screen shots, and explains the results with satisfactory clarity, comprehensiveness, and organization. • Discussions are not comprehensive, and it misses some important points.	• Fail to meet the requirements and incorrect outputs are shown. • The R programme cannot be executed, and incorrect outputs are shown • Minimal or no comments is added to the scripts. • The report includes the screen shots, and unclearly or loosely explains the results. • Discussions are not well focused, and it misses the important points.	• No submission or late submission.
Question 2 (40%)				
Question 3 (40%)				

~~END OF ASSIGNMENT 01~~