Assignment 03 Guidelines

Lauren Kennedy

20/09/2021

The overall task

You are to pretend that you work for a company that streams movies and TV shows. You have been given two different datasets that the company has. Your aim is to produce a report that describes recommendations for the company that you can make from the data you have been given, as well as recommendations for the future. You should aim for transparency in how you report what you did. Be careful to say what you did and why you did it.

Working in pairs

You have the option to complete your assignment as a pair or individually. If you choose to work in a pair you will both receive the same grade. You will need to make the decision to work as an individual or as a pair by **October 5th, 2021**, which you will do by allocating yourself to a group as either an individual or with one other person on Moodle. If you work as a pair you will only submit one assignment, but both of you will need to approve the submission. I have opened a forum specifically for finding a pair to work with. You must choose someone from within your unit code.

The two datasets

For this assessment you will have two related datasets to work with. The first dataset is in `netflix_metadata`. It comes from a TidyTuesday dataset from the 20th of April this year. I have modified it slightly to have the following columns:

- `show_id1` An id that is unique to that particular show or movie
- `title` The title of the show or movie
- `length` A numeric value that contains the length in minutes (if movie) or seasons (if tv show)
- `n_listed_actors` The number of actors listed
- `n_listed_genre` The number of genres the product is listed under
- `min_age_rating` A numeric value that corresponds to the minimum age that the show is recommended for, including products that are recommended for parental guidance.
- `united_states` Whether the product was created in the US (1) or not
- `movie` Whether the product is a movie (1) or TV show (0)
- `drama` Whether the show is classed in the drama genre (1) or not
- `older` Whether the show was released more than 5 years after it was created.

The second dataset is in `user_data`. This is a tiny subset of the Netflix Prize data which I have cleaned and manipulated for you. It has 52 of the most common movies from the Netflix Prize data and ratings from 150 users. This data is distributed for educational purposes only and not intended for redistribution or use beyond this class. If you are interested in the original data you can find it on Kaggle (https://www.kaggle.com/netflix-inc/netflix-prize-data/tasks). This dataset contains the following variables:

- `movie_id` A unique numeric id for the the movie

- `year` The year of release
- `name` The name of the movie
- `user_id` The id of the user
- `rating` The rating from the user

You will not need to join these two datasets together in this assignment.

The task

You will use these two datasets to make recommendations for a streaming service. We would like you to complete these recommendations in three parts, outlined below.

Word count

You have 2000 words total (+10%) to complete this assignment. You can allocate these words how you wish but I suggest Part A - 1000, Part B - 750, Part C - 300

PART A: Netflix metadata (25 marks)

Using the Netflix metadata, create two different possible sets of clusters (you may consider a subset of this data) using the methods we discussed in class. Write a short report including one visualization only. You should cover the following:

1. How you created the two cluster sets (including choice of number of clusters, which variables you included, methodological decisions etc.), and why you think comparing between the two would be interesting. (5 marks)
2. The different profiles created by the clusters and the similarities/differences (5 marks)
3. Discuss the stability of the clusters by splitting your data into two random datasets and running the algorithm separately on both (5 marks)
4. Which method you would use in practice and why (5 marks)
5. Recommendations for how you would suggest using what you learned from this analysis in the company. (5 marks)

Part B: User data (25 marks)

Now the company has presented you with a small subset of user data. Use this data to make recommendations of movies for users. Write a short report including one visualization only. You should cover the following:

1. The analysis your ran, and the decisions you made (including the benefits/disadvantages of these decisions) (5 marks)
2. Choose a recommendation rule that you think the company would use. Use it to explain concepts like support, lift and confidence in words that could be used with a non-technical audience. (5 marks)
3. Consider the user ratings. For the rule from 2) how often does the user not like the consequent using your method? Is your rule useful when this is considered? (5 marks)
4. Make recommendations for how you would suggest using what you learned from this analysis in the company. (5 marks)
5. Considering the users perspective, what benefits and disadvantages are there if the company decided to expand this sort of analysis (5 marks)

Question C: Connecting A and B (15 marks)

Using two sets of data focused on Netflix offerings, you have completed two very different types of analysis. Discuss the following:

1. If the company had to focus on just one analysis method, which would you recommend they focus on and why? (5 marks)
2. If the company wanted to connect the two styles of analysis, how would you suggest they do this? (5 marks)
3. If the company wanted to invest additional data collection, what additional data would you recommend they collect, and what would you suggest they do with it (5 marks)

Code (33 marks)

- Code should be effective at running the analysis described in the text (30 marks) 5 marks are assigned for successful completion of code for each of the first 3 dot points for A and B.
- Code is clearly written, completed in full in the RMD and does not include unnecessary code (for example, package installation or use of the View or str command is best completed in the console) (2 marks)
- The Rmd should knit for us assuming we were to place it in the analysis folder of the Rmd project. (1 mark)

Overall report (12 marks)

- The writing used throughout the document is clear and uses correct spelling and grammar (2 marks)
- For each figure (5 marks)
  - Figures and visualizations should be used effectively and be clearly presented (1 mark)
  - Figures should include a short caption (excluded from word count) of 1-2 sentences that describes the plot and how it can be used (2 marks)
  - The figure should be appropriately referenced within the report (1 mark)
  - The figure should have appropriate and legible axis labels and text (1 mark)

Marking Rubric

Each dot point in A-C will be assessed using the following rubric

- 0 Point was not covered
- 1-2 Either large errors of understanding were made when covering the point, or the point was covered very superficially with little justification
- 3-4 The decision (recommendation/method/etc) was clearly stated, and the justification is clear and correct but not specific to the specific application.
- 5 The decision (recommendation/method/etc) was clearly stated, and the justification is clear, correct and demonstrates an application of core concepts to the specific application challenges

Each of the first three dot points in sections A and B require significant code and analysis. Marks will be awarded according to this rubric for each point

- 0 There was no attempt this or the attempt did not produce viable results
- 1-2 There were large errors or this the code was incomplete

- 3-4 The code is mostly correct, or entirely correct but very complex relative to the task (e.g., using 10 separate lines to take the mean of 10 groups when group_by and summarise could have been used)
- 5 Code is clear, concise and correct