Final Project: Guidance Document

Author Name

Due Monday, May 2, 2022 by 6:00pm EST

Purpose

General: This document provides instructions on what each student should strive to complete for the final project. Each student also needs to create a Git Hub repo titled 'STAT184-SP22-Project' and invite their instructor to their repo using your instructor's Git Hub name 'AndyWiesner'. During the course of completing this project, each student must make at least FOUR commits/pushes to their repo.

Final Project Requirements (all chunks must include relevant level-3 headers)

About the data

Summary: This data describes 5-star rating and free-text tagging activity from MovieLens, a movie recommendation service. It contains 100836 ratings and 3683 tag applications across 9742 movies. These data were created by 610 users between March 29, 1996 and September 24, 2018. These data were generated on September 26, 2018. Users were selected at random for inclusion. All selected users had rated at least 20 movies. No demographic information is included. Each user is represented by an id, and no other information is provided. The data are contained in the files `movies.csv`, `ratings.csv` and `tags.csv`. More details about the contents and use of all these files follows.

- (movies.csv) Only movies with at least one rating or tag are included in the data set. These movie ids are consistent with those used on the MovieLens web site. Movie ids are consistent between `ratings.csv`, `tags.csv`, and `movies.csv`(i.e., the same id refers to the same movie across these three data files)
- (ratings.csv) All ratings are contained in the file `ratings.csv`. Each line of this file after the header row represents one rating of one movie by one user, and has the following format:

  userId,movieId,rating,timestamp

  The lines within this file are ordered first by userId, then, within user, by movieId.

  Ratings are made on a 5-star scale, with half-star increments (0.5 stars - 5.0 stars).

  Timestamp represents seconds since midnight Coordinated Universal Time (UTC) of January 1, 1970.
- (tags.csv) All tags are contained in the file `tags.csv`. Each line of this file after the header row represents one tag applied to one movie by one user, and has the following format:

  userId,movieId,tag,timestamp

  The lines within this file are ordered first by userId, then, within user, by movieId.

  Tags are user-generated metadata about movies. Each tag is typically a single word or short phrase. The meaning, value, and purpose of a particular tag is determined by each user.

  Timestamp represents seconds since midnight Coordinated Universal Time (UTC) of January 1, 1970.

Packages and Data

First chunk will load the only two packages to be used for this project 'tidyverse' and 'lubridate'. The chunk will also read in the three data sets into object names related to that data file: 'Movies', 'Ratings', and 'Tags'

Data Wrangling (Initial)

Description: You will use three chunks to complete the following wrangling tasks.

(1) Convert the timestamp data to the actual date in the ratings and tags files. These dates should appear in new variable column 'rating_date' and 'tag_date' in their respective data files. The final size of both files should be 100836 obs and 5 variables. Publish in notebook the first 3 observations for each file which will verify you properly converted the timestamp data to date.

(2) Create a new data object that joins data where person both rated AND tagged movie. Then join movie info. The final size of this file should be 3476 obs and 7 variables with the 7 variables (in order) of: 'userId', 'rating', rating_date', 'tag', 'tag_date', 'title', 'genres'

Publish in notebook the first 3 observations of this data file.

(3) Create a new data object that includes the data from file created in previous chunk and the new variable 'release_year' based on year in () end of title. Be sure remove year from title. Basically, move the year in the title variable to the new variable 'release_year' but without parentheses. This new data object should have 3476 obs and 8 variable. Publish in notebook the first 3 observations of this data file.

Data Visualization

Answer the following 2 questions by creating a graph that answers that question and include a meaningful, interpretive statement of that graph as it relates to the question. Use one chunk per question where the chunk will both manipulate and graph the data.

(1) What were the top 10 rated movie genres? That is, what were the ten genres rated most frequently?

(2) *Your question* The student should develop their own question and create a graph that answers that question plus meaninful, interpetive statement. The graph should include at least two variables, and use either color or fill options.

Other requirements (Nothing for you to report in this Guidance Document)

(A) *Code quality:* Code formatting is consistent with Style Guide Appendix of DataComputing eBook. Specifically, all code chunks demonstrate proficiency with (1) meaningful object names (2) proper use of white space especially with respect to infix operators, chain operators, commas, brackets/parens, etc (3) use of `<-` assignment operator throughout (4) use of meaningful comments.

(B) *Overall Quality:* Submitted project shows significant effort to produce a high-quality and thoughtful analysis that showcases STAT 184 skills. (2) The project must be self-contained, such that the analysis can be entirely rerun without errors. (3) Analysis is coherent, well-organized, and free of extraneous content such as data dumps, unrelated graphs, and other content that is not overtly connected to the research question.

(C) *Git Hub Repo* The student's Git Hub 'STAT184-SP22-Project' repo will demonstrate at least four pushes/commits to this repo. Each Student must invite their instructor to their repo using your instructor's Git Hub name 'AndyWiesner'