Introduction to Data Science, WMCS16002, semester 1a 2017

1 Homework

Due September 10, 2017 11:59:00 (midnight) CET
Submit by creating a pull request on GitHub

Your submission should consist of

- A brief written report (preferably PDF (latex template available on Nestor) or dynamic report formats (RMarkdown, etc.)
- Source code files which generate the solution, tables, visualizations used in the report. The code can contain much more than what is finally used in the report — please use comments to structure the source code and provide a README.txt file informing how to run it.

Note that you are free to use whatever programming language you want if not stated otherwise in the assignment. Be aware that the difficulty of an exercise might scale differently dependent on what language you use!

You have been allocated to work in groups of three to four people to mix different backgrounds. This is an interdisciplinary course therefore we suggest you take advantage of your complementary background (including Mathematics, Astronomy, Engineering and Computing Science). You can indicate in the last section in case you did not contribute equally and there will be a peer evaluation in the end where you can compliment good teamwork and the opposite. If each of you use the GitHub rigorously we could in principle follow your actions in every detail in case of trouble.

Remember that you have to pass every homework assignment (but one) to pass the course. Furthermore, note that **plagiarism is fraud** and we take it serious. If we find it in your submissions you risk being expelled.

Submission

- The bonus parts are still sent with the pull request, but remember to report on contributors.
- Make a new folder for every assignment (this is the first one). Every assignment's folder should have clear instructions on how to run your code.
- Do not commit massive data files to the repository, leave clear instructions on what and where.

Good luck and have fun working on the exercises!

Before Anything Else

a) Read up on git and GitHub[1] if you are not already familiar with distributed version control. We organize a GitHub tutorial in the first practical on Friday in case you are interested to learn more about it.
b) Create a GitHub account.
c) Fill in the form at `https://goo.gl/forms/CtbEQYKASlVtzcoI2` (the one sent via email) using a student.rug.nl or rug.nl email account so the TAs can add you to your team repository.
d) Join the repository of your team (`https://github.com/RUG-IDS/team-xx`).
e) Clone your repository's `dev` branch and prepare your assignments here.
`git clone -b dev git@github.com:RUG-IDS/team-xx.git`[2]

Note: Only **we** have push access to the `master` branch. You work on the `dev` branch. Each of the assignments should be in a separate folder. When ready, you submit your work by sending a pull request to the `master` branch of your team's repository. We assess the work that is submitted through the pull request.

1.1 Identify Data Types (10P)

Classify the following attributes as binary, discrete, or continuous. Also classify them as qualitative (nominal or ordinal) or quantitative (interval or ratio). Some cases may have more than one interpretation (as the eyecolor example from the lecture as color or amount of pigment ;-)), so briefly indicate your reasoning if you think there may be some ambiguity.
Example: Age in years. *Answer*: Discrete, quantitative, ratio

a) Brightness as measured by a light meter
b) Brightness as measured by people's judgments
c) Time in terms of AM or PM
d) Coat check number (certain places offer you to leave your coat to someone who, in turn, gives you a number tag that you need to claim it back when you leave)

Hollywood Data Science

You might have encountered the Internet Movie Database IMDb `http://www.imdb.com`, which is a collection of film related information. Most of the data can be downloaded as plain text files and some tools are available allowing to search and display information.

Download the file `movievalue.csv` we provided in Nestor. This file contains limited information about the movies. To obtain more information there are two ways:

FTP The FTP server of Freie Universität Berlin (Germany) contains a subset of IMDb data of manageable size. From `ftp://ftp.fu-berlin.de/pub/misc/movies/database/` We also provide an `imdb-data-parser-master.tar.gz` which you might use for the data import if you like.

[1] A good source is `https://help.github.com/articles/about-pull-requests/`
[2] Or if you want to use HTTPS: `git clone -b dev https://github.com/RUG-IDS/team-xx.git`.

API The data can be obtained also from the OMDB API (`http://www.omdbapi.com/`, you can use this key: 863c5282) or TMDB API (`https://developers.themoviedb.org/3/search/search-movies`, key: 3ce50e41bbff335a1a1a7a054f2b141b). However, there you have to request information for every movie individually.

1.2 Collect It ... Link It! (50P)

Collect and clean the data about the movies in the `movievalue.csv`. Enrich that data collection by acquiring also the `Genre, imdbRating, imdbVotes` (and optional also Director, Country, PG rating, etc.) for at least 1000 of the given movies for further analysis.

TIP: One of the two ways of getting the data (FTP vs. API) is easier than the other. You have to figure out which.

If you are using R, Matlab or Python, the following might be useful:

a) `data.frame` (R)
b) `table` (Matlab from R2013b)
c) `DataFrame` (Python).

Note, linking the data might be tricky since names might not be unique (episodes of series often reuse names of movies and pre/sequels may be written differently)!

1.3 Think About Types (20P):

The data set(s) you constructed in 1.2 contains features like "ReleaseDate", "Movie", "Production Budget", "Production company","Domestic Gross", "Worldwide Gross", "genre", "imdb rating", "number of imdb votes", "director", etc.
Determine the data type of each of them and explain your decision shortly.

1.4 And Finally ... Analyze It! (20P):

a) Perform exploratory analysis on the data set and present some of your observations.
b) Select three descriptive questions on the data set and present their answers using diverse approaches if needed (e.g. table, scatter plot, box plot and/or histogram).

1.5 Bonus (+10P):

Analyze at least 2 of the following 4 questions:

- Acquire ratings from `www.rottentomatoes.com` and compare them with the IMDB ratings. What do you observe? How and why are they different?
- Do you find evidence that the involvement of a certain director or production house in a film leads to better ratings or box office success than others? Explain your findings

- Find out if certain actors appear more often than others in certain genre(s) of movie? Or, are certain genre(s) of movie more associated with certain production houses (example, Pixar with animation, etc).
- You've probably noticed that some movies have missing data or some movies couldn't be found at all. We can treat the missingness of data as information itself. Perform exploratory analysis that uses missing data as a data type (e.g. plotting release date versus frequency of missingness).