Introduction to Data Science, WMCS16002, semester 1a 2018

1 Homework

Due September 7, 2018 8:00 o'clock CET
Submit in Nestor in Content "First week: individual assignment"

The first assignment is the only individual homework in this course.

Note that participation in it is **mandatory**, since we will use the participation in the creation of the groups together with the questionnaire. If you neither submit a homework nor provide answers to the questionnaire at the end of the first week we assume that you are a ghost enrollment who does not actively participate and we will not assign you to a group. Therefore you would not be able to participate in the future assignments!

Good luck and have fun working on the exercises!

Before Anything Else

a) Fill the questionnaire we will use to determine the group memberships. Find it by following the link https://goo.gl/forms/4fx7gNmvzzKkMPey1 or scanning the QR code on the right side. For more information read the entry in Nestor found under "Information".
b) Read up on git and GitHub[a] if you are not already familiar with distributed version control. We organize a **GitHub tutorial in the first practical on Tuesday** in case you are interested to learn more.
c) Create a GitHub account.

[a] A good source is https://help.github.com/articles/about-pull-requests

1.1 Digital Test (100P)

Submit the answers in the digital test found in Nestor named **First week: individual assignment** under "Content".
The following sections give you an impression how the group assignments will look like during the rest of the course. We highly encourage you to perform the analysis to practise the material.

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.
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!

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.

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:

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.

a) Determine the data type of each of them and explain your decision shortly.
b) Argue shortly what preprocessing would be useful for at least 3 example variables.

1.4 And Finally ... Analyze It!

a) 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).
b) Perform exploratory analysis on the data set and present some of your observations.

1.5 Bonus:

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).