Description of the project to be submitted

Stephan.Huber@hs-fresenius.de

HS-Fresenius: Data Science for Business

Contents

Word count: 1586

Abstract

In the following I describe the project that needs to be submitted in the course *Data Science for Businesses.* I give some hints for your efficient progress and success, I introduce the elements and files to be submitted, and I describe how I evaluate the submissions.

1 Main goal

Course description *''Students complete this module with a project work. The project work includes a project report (15-20 pages) and a project presentation (20-30 minutes).''*

Project work Create a *swirl* learning module, write a report of the project, and present the content of the *swirl* learning module. As a final product, you can publish your swirl learning modules on Github or elsewhere to show your R proficiency.

2 Details

2.1 The learning module

- The content of the learning module should be strongly related to the content of the book *R for Data Science* by Wickham and Grolemund (2018).

- Each student is assigned some functions of R with a focus on functions of the tidyverse package. Each student should cover the content of a particular section of the book of Wickham and Grolemund (2018).

- The learning module should show how the respective functions can be used in a data analytical investigation. Students are totaly free in choosing certain steps of a data analytical procedure. However, I strongly encourage students to discuss their ideas.

- The processing time to complete a learning module should be approximately 30 minutes.

- The module should be written using the swirlify package. For more complete instructions

about writing swirl courses please see the swirlify documentation.

- In terms of content, there are only a few specifications:
  - The learning module can build on concepts covered in previous parts of the book.
  - Try to use check questions to test the content introduced in the module. This should ensure the understanding of the students.
  - Since the goal is to create a module that can be published, be aware of copyright issues! Do not copy and use content from others without mentioning it and without being allowed to do so on their part. In data science and the open source community, many people use Creative Commons Copyright Licenses. Depending on the particular license, it may be allowed to distribute, remix, adapt, and build upon the work of others, even commercially.
  - **Use World Bank data** to address the capabilities of the respective R functions.
    * Therefore, I recommend using the *WDI* package. It allows to search and download data from over 40 datasets hosted by the World Bank. More information on this package can be found in the documentation of the package.
    * Try to select variables that are interesting and easy to understand.
    * Make sure that the variables used are appropriate to represent the properties of the particular R functions being explained.

2.2   The report

- The report should be about 4000-6000 words (which is about 15 double-spaced pages).

- Unlike an *academic paper*, this is a report in which you should just document, discuss, and present your project. The report should introduce your work to me. It's similar to reports that you'll have to write in business, where your boss wants to know

  - What you did,
  - why you did it the way you did,

- – what obstacles you overcame,

- – what challenges, problems and weaknesses remain, and

- – how you would suggest proceeding with your work if you would have more time and ressources avalailable.

- While graphical representations often help improve the communication of sophisticated content, it's not necessary. In other words, don't try to impress with a fancy layout or anything like that. Focus on the content and getting your knowledge across to the reader! Anything that helps with this is welcome.

- Guide and motivate the reader and outline the target audience of your work. Usually, the introduction is a good place to introduce the scope and content. In particular, make clear what is found in each section.

- Be concise. Remove all unnecessary repetition. Read each sentence several times and ask yourself if it is concise and clear and if it fits with what was said before and after.

- You don't have to explain everything from scratch, but if you shorten things that are essential to understanding, you should give the reader a source where they can find the missing pieces and teach themselves.

- The paper should be written with **R Markdown**. The .Rmd file with which this document is written, can be used as a template. This file is hosted at my GitHub page: Information about writing and publishing with R Markdown can be found here:

  - – R Markdown from R Studio
  - – R Markdown: The Definitive Guide written by Xie, Allaire, and Grolemund (2018)

- Insert R code in your R markdown file by typing the chunk delimiters (see the keyboard shortcut *Ctrl+Alt+I* for Windows and Linux based OS and *Cmd+Option+I* for Macintosh OS) or this lesson).

- The outline of the paper should contain the following building blocks:

- Title and all common details (name, email, . . .).

- Word count (see this rmd file on how to give a word count).

- Abstract of the paper (which highlights the content of the document).

- A summary of the Swirl learning module so that students can decide in advance whether to take the module or not.

- The Swirl learning module in a printed version (see below how to add it to the rmd file)

- A glossary with functions used in the swirl module.

- Further descriptions to the project, which may include, for example:

    * How the datasets were prepared for the swirl module.

    * How to access and install the module (with the .swc file, online via GitHub, . . .).

    * Challenges you had to overcome when creating the module (keep it short!).

    * Explanations and comments on the didactic concept of the module.

    * . . .

2.3 The presentation

- Write the presentation using R Markdown and publish it as .html and/or a .pdf file.

- Focus on the important things.

- Try to stay on schedule.

- Nobody is perfect and the project is done under time pressure. So don't try to sell yourself too hard. If you see weaknesses in your work, this is a good place to discuss them.

3 Submission

- Submission deadline for academic papers and written assessments: **25 Jul 2022** *(please verify! I am not a member of the examination office and I have not studied the specific rules that are valid for you!)*

- Upload only **one .zip file** that contains:

  1. the paper as a (1) .pdf, (2).html file and (3) .Rmd file

  2. the swirl course as a (4) .swc file,

  3. additional files, if necessary for me to evaluate your work.

4 Evaluation scheme

- *40 % – Quality and execution of the swirl module*

  - Does it work technically?

  - Is the length appropriate?

  - Does it make sense didactically?

  - Is it motivating to learn with your module?

  - Is the module interactive and are the type of question appropriate?

  - Are content and R functions explained well?

  - Does it relate to the book of Wickham and Grolemund (2018)?

  - Are the sample datasets well chosen and introduced?

  - . . .

- *30 % – Quality and execution of the paper*

  - Is it well-written?

  - Are formal criteria OK? (structure, length, format)

  - Does it give a comprehensive overview about the project?

  - . . .

- *30 % – Quality and execution of the presentation*

- Is the presentation well structured?

- Is it easy to follow and informative?

- How does the presenter answer questions of the audience?

- . . .

5 Projects to be assigned

1. Data Visualization with ggplot2 (Wickham and Grolemund 2018: ch. 1, p. 3-36).

   - max. 2 learning modules that can be worked on in a group of two people.

   - ggplot2 package

2. Data Transformation with dplyr (Wickham and Grolemund 2018: ch. 3, p. 43-76).

   - max. 2 learning modules that can be worked on in a group of two people.

   - dplyr package

3. Exploratory Data Analysis (Wickham and Grolemund 2018: ch. 5, p. 81-110).

   - max. 3 learning modules that can be worked on in a group of three people.

   - dplyr and ggplot2 packages

4. Tibbles with tibble (Wickham and Grolemund 2018: ch. 7, p. 119-124).

   - max. 2 learning modules that can be worked on in a group of two people.

   - tibbles package

5. Data Import with readr (Wickham and Grolemund 2018: ch. 8, p. 125-146).

   - max. 1 learning module.

   - readr package

6. Tidy Data with tidyr (Wickham and Grolemund 2018: ch. 9, p. 147-170).

   - max. 2 learning modules that can be worked on in a group of two people.

   - tidyr package

7. Relational Data with dplyr (Wickham and Grolemund 2018: ch. 10, p. 171-194).

   - max. 2 learning modules that can be worked on in a group of two people.

- tidyr and dplyr packages

8. Strings with stringr (Wickham and Grolemund 2018: ch. 11, pp. 195-222).

    - max. 1 learning module

    - stringr package

9. Factors with forcats (ch. 12, p. 223-236).

    - max. 1 learning module

    - forcats package

10. Dates and Times with lubridate

    - max. 1 learning module

    - lubridate package

```
rmarkdown::render("22-04_dsb-project-description.Rmd", "all")
```

Literature

Wickham, Hadley, and Garrett Grolemund. 2018. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data.* Sebastopol, CA: O'Reilly.

Xie, Yihui, Joseph J Allaire, and Garrett Grolemund. 2018. *R Markdown: The Definitive Guide.* Chapman; Hall/CRC.