Merkitse tehdyksi

Assignment 2: Tasks and Instructions

The last week was a warmup, now we're really getting started! Please note that these exercises may require significantly more work from you. The theme for the week is regression analysis. Have fun and don't be afraid to ask for help using the **Moodle Q&A forum**.

After completing all the phases you are ready to submit your Assignment for the review (see below!).

General instructions

Complete the Exercise Set 2 to go through various R tricks and to obtain the R code needed in the Assignment.

From now on, the Assignments consist of **1) Data wrangling exercises** and **2) Analysis exercises**. Both types of exercises consist of smaller parts, and each part is assigned a maximum number of points for completing the part.

During the Data wrangling exercise you will create an analysis dataset for the Analysis exercise. During the Analysis exercise you will explore the data, perform analysis and interpret the results.

The data wrangling exercise is worth a maximum of 5 points and the analysis exercise is worth a maximum of 15 points. **We also provide a direct link to the analysis dataset in the case you get stuck with the data wrangling exercise.**

Remember that your work should appear on GitHub. You can (and should) update your GitHub all the time between your work sessions. Remember that to update your course diary, you need to "knit" your index.Rmd file into an index.html document. Also of course, you always need to push your changes to GitHub.

You will be working on a data set for which you can find more information here.

Data wrangling (max 5 points)

During the data wrangling exercises you will pre-process a data set for further analysis. To complete the data wrangling part, you only need to produce an **R script**, no output in your course diary is needed. Use code comments to make your code easier to read. Always write your name, date and a one sentence file description as a comment on the top of the R script (include a reference to the data source). We recommend using RStudio for writing R code.

1. Create a folder named 'data' in your IODS-project folder. Then create a new R script with RStudio. Write your name, date and a one sentence file description as a comment on the top of the script file. Save the script for example as 'create_learning2014.R' in the 'data' folder. Complete the rest of the steps in that script.
2. Read the full learning2014 data from http://www.helsinki.fi/~kvehkala/JYTmooc/JYTOPKYS3-data.txt into R (the separator is a tab ("\t") and the file includes a header) and explore the structure and dimensions of the data. Write short code comments describing the output of these explorations. **(1 point)**
3. Create an analysis dataset with the variables gender, age, attitude, deep, stra, surf and points by combining questions in the learning2014 data, as defined in the Exercise Set and also on the bottom part of the following page (only the top part of the page is in Finnish). http://www.helsinki.fi/~kvehkala/JYTmooc/JYTOPKYS2-meta.txt. Scale all combination variables to the original scales (by taking the mean). Exclude observations where the exam points variable is zero. (The data should then have 166 observations and 7 variables) **(1 point)**
4. Set the working directory of your R session to the IODS Project folder (study how to do this with RStudio). Save the analysis dataset to the 'data' folder, using for example **write_csv()** function (*readr* package, part of *tidyverse*). You can name the data set for example as learning2014.csv. See **?write_csv** for help or search the web for pointers and examples. Demonstrate that you can also read the data again by using **read_csv()**. (Use `str()` and `head()` to make sure that the structure of the data is correct). **(3 points)**

Analysis (max 15 points)

The Analysis exercise focuses on performing and interpreting regression analysis. For completing the Analysis exercises, include all tl codes, your interpretations and explanations in the R Markdown file chapter2.Rmd, which already exists in your course project folder.

If you wish, you can knit the chapter2.Rmd file as a html document anytime to see how it looks. But for submission, the output of your Analysis should appear in your course diary. For this **you need to update your local index.html file by knitting the index.Rmd file** (which includes chapter2.Rmd as a child file) and **then push the changes to GitHub**.

Write a continuous report with a clear structure. There is no need to repeat the assignments in your course diary. The focus of your work should be on the clarity of your report. For full points you should be able to show **an understanding of the methods and results** used in your analysis. Feel free to also use material outside this course as learning sources. **Clear, understandable and comprehensive explanations are worth full points.**

R Markdown Hint: When you knit the document, the working directory is temporarily set to be the folder where your R Markdown file is located. This is good to be aware of when reading data for example.

1. Read the students2014 data into R either from your local folder (if you completed the Data wrangling part) or from this url: https://raw.githubusercontent.com/KimmoVehkalahti/Helsinki-Open-Data-Science/master/datasets/learning2014.txt ➚ . (The separator is a comma "," and the file includes a header). Explore the structure and the dimensions of the data and describe the dataset briefly, assuming the reader has no previous knowledge of it. There is information related to the data here. ➚ **(0-2 points)**

2. Show a graphical overview of the data and show summaries of the variables in the data. Describe and interpret the outputs, commenting on the distributions of the variables and the relationships between them. **(0-3 points)**

3. Choose three variables as explanatory variables and fit a regression model where exam points is the target (dependent, outcome) variable. Show a summary of the fitted model and comment and interpret the results. Explain and interpret the statistical test related to the model parameters. If an explanatory variable in your model does not have a statistically significant relationship with the target variable, remove the variable from the model and fit the model again without it. **(0-4 points)**

4. Using a summary of your fitted model, explain the relationship between the chosen explanatory variables and the target variable (interpret the model parameters). Explain and interpret the multiple R-squared of the model. **(0-3 points)**

5. Produce the following diagnostic plots: Residuals vs Fitted values, Normal QQ-plot and Residuals vs Leverage. Explain the assumptions of the model and interpret the validity of those assumptions based on the diagnostic plots. **(0-3 points)**

After completing all the phases above you are ready to submit your Assignment for the review (using the Moodle Workshop below). Have the two links (your GitHub repository and your course diary) ready!

Viimeksi muutettu: torstaina, 17. marraskuuta 2022, 11.16