Data Analytics in R
Session 7

Maria Kunevich

Homework feedback

- Comments to the submitted gists are below the gist
- Solutions will be uploaded to GitHub in the Assignments folder
- Let’s have a look at some common errors (R scripts)
- Commenting is important !
- Feedback on HW1
- Feedback on HW2 (task with adding columns and rows)
- Feedback on HW3 (the difference between ‘table’ vs ‘data.table’)
- Feedback on HW4

Homework assignment 4/5

- HW4 will be combined with HW5, the total number of points is 10
- Submitting the html file that you ‘knit’ from your Rmarkdown file
- Today we’ll go over the steps of using RMarkdown and creating html files as an output
- How I’ll check the homework: you send me a link to your .html file through your gist or directly to my email
- I’ll upload the file (yourlastnameHW4) to my Github Reporting repository and check if I can render your code on my webpage

Homework feedback

EDA - Exploratory Data Analysis

Resource: *R for Data Science*

https://r4ds.had.co.nz/exploratory-data-analysis.html#exploratory-data-analysis

Exploring your data in a systematic way:

- ❏ Generate questions about your data
- ❏ Visualise and transform your data to find answers to your questions
- ❏ Redefine your questions or create new ones
- ➢ You see the quality of your data and can say if expectations are met
- ➢ You deploy EDA tools (visualisation, transformation) to do data cleaning

Homework feedback

Main aims for HW 4

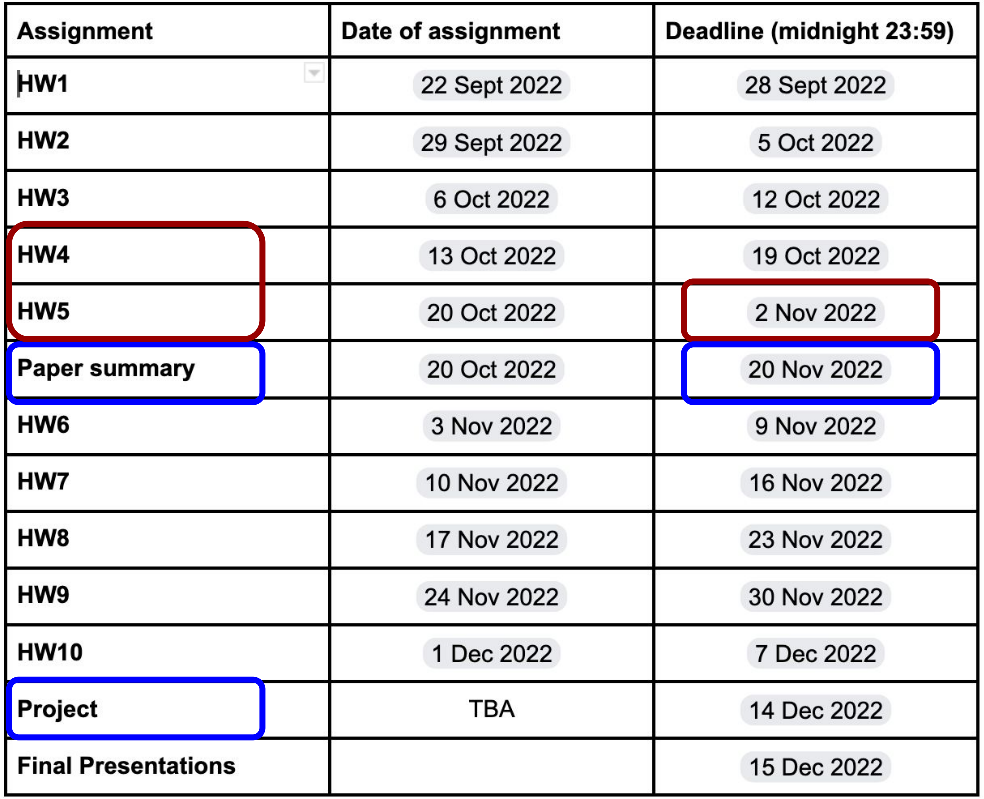
- To understand how to load data into R from external resources
- To understand how to perform EDA as the first step of data analysis, what kind of questions to ask during this stage
- To understand how to work with RMarkdown and render the output to .html file
- To understand how to create basic plots and include them in a report
- To learn how to communicate the results of EDA and first stages of data analysis through writing a short report

Next seven sessions

Providing tools and information to prepare you to work on your projects

✓ Find data and import it into R, describe your dataset
✓ Formulate questions for the first stage of EDA
✓ Visualise your data (‘ggplot2’)
✓ Report EDA results
❏ Data cleaning (‘tidyr’)
❏ Data wrangling (‘dplyr’)
❏ Exploring variation, covariation, covariance and correlation
❏ Hypothesis testing (inferential statistics)
❏ Data modeling (regression models)
❏ Communicating your results

Assignments deadlines



Assignment	Date of assignment	Deadline (midnight 23:59)
HW1	22 Sept 2022	28 Sept 2022
HW2	29 Sept 2022	5 Oct 2022
HW3	6 Oct 2022	12 Oct 2022
HW4	13 Oct 2022	19 Oct 2022
HW5	20 Oct 2022	2 Nov 2022
Paper summary	20 Oct 2022	20 Nov 2022
HW6	3 Nov 2022	9 Nov 2022
HW7	10 Nov 2022	16 Nov 2022
HW8	17 Nov 2022	23 Nov 2022
HW9	24 Nov 2022	30 Nov 2022
HW10	1 Dec 2022	7 Dec 2022
Project	TBA	14 Dec 2022
Final Presentations		15 Dec 2022

Paper summary

Objectives:

- learn how to analyse and provide an in-depth report on someone's work
- learn how to find scientific papers related to your interests
- help you improve your skills in analysing complex text and summarising it in a short summary
- practice your writing skills
- introduce me to lots of interesting research through reading your summaries :)

Paper summary

- Resources: good guidelines:
- http://courses.washington.edu/ordinary/summary.pdf
- https://writingcenter.uconn.edu/wp-content/uploads/sites/593/2014/06/How_to_Summarize_a_Research_Article1.pdf
- First step: look for the articles that use **R as a tool for analysis** in your field
- Resources: GoogleScholar:
- https://scholar.google.com/schhp?hl=en&as_sdt=0,5
- Access through TLU library to databases with published research:
- https://login.ezproxy.tlu.ee/login

Paper summary

❑ Deadlines: **2 Nov 22**

- On our Miro board: provide a link to your article and a brief description why it’s relevant for the course
- https://miro.com/welcomeonboard/SUw3RHpXWjBpUDF2V0dwTWhkOFVVUnlUTnZ4Qm1oVGtSTVN0SmNCUmJyODhydU9hUzA3VUpNZVZHRnNBenhqVHwzMDc0NDU3MzYzNjI1MDIzMjY2fDI=?share_link_id=590330886950
- Guidelines: → the paper is published in a respectful source
  → the paper uses R for analysis
  → it’s not a guide or a tutorial or a description of package

Project: stages

❑ Deadlines: **2 Nov 22**

- On our Miro board: create a team of 2 students, provide basic description of your project, provide a plan for 2 weeks' intervals of what you expect to complete during this time
- **Stage 1: 16 Nov** (finding datasets, first EDA)
- **Stage 2: 30 Nov** (data cleaning and wrangling, first hypothesis testing)
- **Stage 3: 7 Dec** (hypothesis testing, data modeling, communication your results)
- Final reports in your repositories: **14 Dec**

Choosing a Project

❑ Align the project with your interests

- **Option 1:** analysis of data you already have or planning to analyse for your bachelor/master thesis
- **Option 2:** reproducing results from the published research
  https://zenodo.org
  https://plos.org/open-science/open-data/
- **Option 3:** creating your own project from the available datasets
  Datasets can be found here: https://avaandmed.eesti.ee/
  https://data.unicef.org/resources/resource-type/datasets/
  http://openclimatedata.net/
  https://data.worldbank.org/

R Markdown

- What is Markdown?
- Markdown syntax: R Markdown cheatsheet
- Inline formatting: *italics*, **bold**
- Inline code: with three backticks ``` my code```
- Block-level elements: headers, list items, blockquotes
- Math expressions

Experiment with the text input in R markdown to change the formatting

R Markdown

- R Markdown and R code chunks
- Chunk options: https://yihui.org/knitr/options/
- https://rmarkdown.rstudio.com/lesson-3.html
- ```{r, my-chunk, echo=FALSE, fig.height=4, dev='jpeg'}```

- Figures and tables
- Caching
- Global options

R Notebooks: https://rmarkdown.rstudio.com/lesson-10.html