**------------------------------------------------------**

**Georgetown University**

**Data Science Basics in R**

Summer 2024

M-F 1:00 PM - 3:30 PM

June 3-7th, 2024

Car Barn 303

**------------------------------------------------------**

**Professor:** Steph Eaneff

Email: [sde31@georgetown.edu](mailto:sde31@georgetown.edu)

Office Hours: immediately after class or by appointment

**Course goals**

This course will build literacy and basic proficiency in statistical programming, with a focus on the skills needed to conduct data analyses in professional healthcare and public health workplaces. We will cover the basics of data management, data cleaning, data visualization, and basic statistical calculations in R, and version control in github. Students will leave with a small portfolio of relevant data visualizations and analyses completed using a real-world public health dataset.

At the end of this course, students should be able to:

* Define statistical programming and identify commonly used tools (e.g., R, Python, SAS)
* Describe the value of version-control and develop basic proficiency with github
* Explain the basics of exploratory data analysis and describe commonly used methods
* Demonstrate an ability to read and write documentation about both data and code, including understanding how to comment code and how to read/write a data dictionary
* Write basic R code, including to do basic arithmetic, read in a dataset, and calculate summary statistics
* Generate data visualizations with R using both base R and ggplot using real-world public health and health diplomacy datasets

**Prerequisites**

The course is open to anyone with an interest in learning the basics of statistical programming. No prior coding experience is necessary, and, typically the majority of students have not used R before. The majority of the datasets we use for worked course examples will be focused on public health and health policy, but the class will be primarily be focused on learning statistical programming – we will explore and learn from these datasets together! No prior course work in public health is required.

**What to bring to class**

If possible, please bring a fully charged laptop (with a charging cord) to class each day. I understand that this might not be possible for everyone, so please don’t worry if you are not able to do this. In the case that you don’t have access to a laptop during class, you will be asked to follow course materials with the professor, with a teaching assistant, or with another student.

**How this course is graded**

*75% Attendance and participation***.** If you are taking this course for credit, you are expectedto attend class and participate in class discussions. Participation will be assessed based on:

* respectful and attentive listening
* asking questions and asking for help when you are stuck
* sharing worked examples during in-class exercises
* minimizing distractions (e.g., checking email, texting)

I sincerely hope you are able to attend every class during the week-long course period, but I also understand that life happens. If you do need to miss a class, please reach out to me directly so that we can make a plan to make up what you’ve missed.

*25% In-class exercises.*During each class, we will work through hands-on exercises together. Students are expected to participate in these exercises and complete them to the best of their ability, including asking questions and asking for help when they get stuck. Exercises will be reviewed and discussed as a group during the class.

No take home assignments will be given. Optional exercises may be provided for additional practice – but they really are *optional* and will not be taken into account as part of grading.

I use the following grading cut-offs: >94%=A; 90-93.9%=A-; 87-89.9%=B+; 84-86.9=B; 80-83.9=B-; 70-80 Cs…

**Course schedule**

|  |  |
| --- | --- |
| **Day** | **Topic** |
| 1 | Introduction to statistical programming |
| 2 | Data management and version control |
| 3 | Exploratory data analysis |
| 4 | Designing data visualizations |
| 5 | Build an online portfolio on github |

During the course, all materials will be available on github at <https://github.com/seaneff/data-science-basics-2024>, including daily schedules and additional details on detailed learning objectives.

**Optional reading**

All course materials are optional to read and review, and are available for free online.

*R basics*

* Grolemund, Garrett, and Hadley Wickham. R for Data Science. <https://r4ds.had.co.nz/>
* Douglas, Alex; Roos, Deon; Mancini, Francesca; Couto, Ana Couto, and Lusseau, David. An Introduction to R. <https://intro2r.com/>

*Version control and github*

* Github Docs. Hello World. <https://docs.github.com/en/get-started/start-your-journey/hello-world>
* The Coding Train, Youtube. Github for Poets. <https://www.youtube.com/watch?v=BCQHnlnPusY&list=PLRqwX-V7Uu6ZF9C0YMKuns9sLDzK6zoiV&index=2>
* (more advanced) Bryan, Jennifer. Happy git with R. <https://happygitwithr.com/>

*Exploratory data analysis*

* National Institute for Technology and Standards (NIST). Engineering Statistics Handbook. Exploratory Data Analysis. <https://www.itl.nist.gov/div898/handbook/eda/eda.htm>
* Childs, Dylan Z. Introduction to Exploratory Data Analysis with R. <https://dzchilds.github.io/eda-for-bio/>
* Fox MP, Murray EJ, Lesko CR, Sealy-Jefferson S. On the need to revitalize descriptive epidemiology. American Journal of Epidemiology. 2022 Jul;191(7):1174-9. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9383568/

*Data visualization*

* Cleveland WS, McGill R. Graphical perception and graphical methods for analyzing scientific data. Science. 1985 Aug 30;229(4716):828-33.
* Journal of the Data Visualization Society. https://nightingaledvs.com/
* From Data to Viz. <https://www.data-to-viz.com/>
* R Graph Gallery. <https://r-graph-gallery.com/index.html>
* Lisa Charlotte Muth. A detailed guide to colors in data vis style guides. <https://blog.datawrapper.de/colors-for-data-vis-style-guides/>
* Elliot Jay Stocks. Making sense of typographic classifications. <https://fonts.google.com/knowledge/introducing_type/making_sense_of_typographic_classifications>

*If you have any difficulty accessing the readings, please let me know. Additional optional readings will also be highlighted in daily course materials (e.g., in slides, on github). These materials are not necessary to successfully complete the course, and are provided as additional resources to explore if and where students want to learn about a given topic more deeply.*

**Additional Protocols and Resources**

Public Health Protocols:If you are feeling ill, please do not come to class. If you are under isolation or quarantine order, please do not come to class. If you are unable to come to class for these reasons, we will make accommodations. There is never a penalty for missing class for legitimate public health or other medical related reasons.

Academic Resource Center (ARC):The ARC assists in many skills areas necessary for academic success through individual consultations or workshops, accommodations to students with disabilities, and facilities and support services. For more information: <http://academicsupport.georgetown.edu/>

Gender Identity*:* We respect and refer to people using the names and personal pronouns that they share; and we invite and encourage people to share them. Guidelines for respecting gender identity are found here: <https://executivefaculty.georgetown.edu/endorsed-policies/>