Week 1 Overview

Welcome! Make sure you have reviewed the syllabus and the expectations for discussion/participation. This course is meant to introduce you to statistical methods in Data Science and the primary language / product that those methods are performed with: R. While it's just a letter, it is the second language you will learn in this program.

In addition to the text, I will also include some other sites and readings for you, especially those of you who haven't done statistics 1) ever or 2) lately. If you are either of these 2, do not hesitate to ask questions. No one will be anything more than gracious that you were brave enough to ask the things they were too afraid to. As a note, *Discovering Statistics Using R* was written in 2012 – so there may be some syntax issues that come up when following along directly with the text. Focus on the concepts, because syntax, programs, shortcuts for doing code change **overnight** – there is no way a text could keep up with how frequently these open-source programming languages evolve. Ask questions, and again focus on the why of what you are doing.

I will say, with bold characters that **this course will not be easy**. However, as always there is truth to the axiom that the more thought and care you put into it the more you will get out of it. Also, my hope is that your curiosity guides you and is far stronger than your worry that you don't know something. The additional part of that axiom is that the more questions you ask me and each other, the more you will get out of it.

Throughout this week we'll get started with RStudio and Github – the goal is to orient you to these two tools!

Contents of the Week

Overview

Readings, Assignments, and Tasks

Introduction Discussion Post

1.1 Discussion/Participation

1.2 Exercise

Final Project Overview

Objectives

After completing this week, you should be able to:

Install R and R Studio

Navigate the R Studio IDE

Understand GitHub, Git, and Forking/Cloning

Weekly Resources

Comprehensive R Archive Network

R Studio

Sage Publications. (2021). Discovering Statistics Using R

RStudio, PBC. (2021). RStudio Documentation

RStudio, PBC. (2021). R Studio Cheatsheets

Week 1 Readings, Assignments, and Tasks

Here are your tasks for this week:

Read the following:
- *R for Everyone:* Chapters 1 -3
- *Discovering Statistics Using R:* Chapter 1
- Rstudio-ide.pdf
- Getting Started with GitHub Desktop (GitHub)
- Introduction to GitHub (GitHub)

Watch the following:
- Getting Started with GitHub (YouTube, 2017)
- Introduction - GitHub & Git Foundations (YouTube, 2013)
- GitHub and GitHub Desktop (YouTube, 2018)

Complete the following:
- Introduction Discussion Post
- 1.1 Discussion/Participation
- 1.2 Exercise
- Review Final Project Overview to understand requirements for final project

1.1 Discussion/Participation

Here are optional topics for discussion via Teams this week. Remember,

these topics aren't required, but if you are struggling to know what to post about, these can be used to initiate discussion!

1. What is R?
2. What is R Studio used for? How is it different from R?
3. What is GitHub?
4. What are packages?
5. What is the difference between qualitative and quantitative methods?
6. What process(es) should be followed when conducting research?
7. What is a variable?
8. How do you prove out a theory?
9. What are independent vs dependent variables?
10. What are the different types of variables?
11. What is the difference between validity and reliability?
12. What are different methods of research?
13. Why does randomization matter?
14. What is frequency distribution?
15. Describe some of the descriptive statistics available and what they mean
16. What is a z-score?
17. How do you fit statistical models to the data?

1.2 Exercise

1. Install R
    1. https://cran.r-project.org/
    2. Nothing needs to be submitted
2. Install R Studio Desktop (Just the free version!)
    1. https://rstudio.com/products/rstudio/download/
    2. Nothing needs to be submitted, but I would encourage you to look around and test out a few lines of code
3. Complete the following GitHub Tutorials
    1. Hello World (GitHub)
    2. Set up Git (GitHub)
    3. Create a Repo (GitHub)
    4. Fork a Repo (GitHub)
        1. The Bellevue GitHub Repo for this course can be found here: DSC 520 GitHub Repo - this is what you want to fork during the tutorial
        2. You will not be completing any pull requests in this course, but it is good to understand the concepts.
    5. Submit your GitHub URL and a screenshot of your

forked repository.

Submission Instructions

The assignment is due by Sunday, 11:59 p.m. CT.

Final Project Overview

You will be working on a research paper for your final project. This project will include identifying a topic/problem that you want to solve using data science. While the final solution to the problem does not need to be provided via programming – you will be doing some exploratory data analysis, transformations, and summary statistics on the data via R. You are welcome to create a model based on what you have learned in this course to solve the problem, but this is not required. Instead, a recommendation is required for a model or method you would implement to solve the problem. There are 3 steps to this project, with the final deliverable being a formal paper (completed in R Markdown) that outlines the problem, shows the analysis done with the data, and concludes with your recommendation for next steps. Each step provides detailed information that you must include in each phase of the project.

Step 1 - Weeks 8 & 9

You will be working on a research paper for your final project. This project will include identifying a topic/problem that you want to solve using data science. While the final solution to the problem does not need to be provided via programming – you will be doing some exploratory data analysis, transformations, and summary statistics on the data via R. You are welcome to create a model based on what you have learned in this course to solve the problem, but this is not required. Instead, a recommendation is required for a model or method you would implement to solve the problem. There are 3 steps to this project, with the final deliverable being a formal paper (completed in R Markdown) that outlines the problem, shows the analysis done with the data, and concludes with your recommendation for next steps. Each step provides detailed information that you must include in each phase of the project.

- Identify a topic or a problem that you want to research. Provide an introduction that explains the problem statement or topic you are addressing. Why would someone be interested in this? How is it a data science problem?
- Draft 5-10 Research questions that focus on the problem statement/topic.
- Provide a concise explanation of how you plan to address this problem statement.
- Discuss how your proposed approach will address (fully or partially) this problem.
- Do some digging and find at least 3 datasets that you can use to address the issue. (There is not a required number of fields or rows for these datasets)
    - Original source where the data was obtained is cited and, if possible, hyperlinked.
    - Source data is thoroughly explained (i.e. what was the original purpose of the data, when was it collected, how many variables did the original have, explain any peculiarities of the source data such as how missing values are recorded, or how data was imputed, etc.).
- Identify the packages that are needed for your project.
- What types of plots and tables will help you to illustrate the findings to your research questions?
- What do you not know how to do right now that you need to learn to answer your research questions?

You can use the following template for Step 1:

- Introduction
- Research questions
- Approach
- How your approach addresses (fully or partially) the problem.
- Data (Minimum of 3 Datasets - but no requirement on number of fields or rows)
- Required Packages
-
    - Plots and Table Needs
    - Questions for future steps

Step 2 - Week 10

At this point you should have framed your problem/topic, described the data, and how you plan to solve the problem. Now you need to move on to the next step of analyzing and preparing the data.

- Data importing and cleaning steps are explained in the text and follow a logical process. Outline your data preparation and cleansing steps.
- With a clean dataset, show what the final data set looks like. However, do not print off a data frame with 200+ rows; show me the data in the most condensed form possible.
- What do you not know how to do right now that you need to learn to import and cleanup your dataset?

- Discuss how you plan to uncover new information in the data that is not self-evident.
- What are different ways you could look at this data to answer the questions you want to answer?
- Do you plan to slice and dice the data in different ways, create new variables, or join separate data frames to create new summary information? Explain.
- How could you summarize your data to answer key questions?
- What types of plots and tables will help you to illustrate the findings to your questions? Ensure that all graph plots have axis titles, legend if necessary, scales are appropriate, appropriate geoms used, etc.).
- What do you not know how to do right now that you need to learn to answer your questions?
- Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.

Some additional questions you may want to consider asking yourself as you work through this section of the project:

1. What features could you filter on?
2. How could arranging your data in different ways help?
3. Can you reduce your data by selecting only certain variables?
4. Could creating new variables add new insights?
5. Could summary statistics at different categorical levels tell you more?
6. How can you incorporate the pipe (%>%) operator to make your code more efficient?

You can use the following template for Step 2:

- How to import and clean my data
- What does the final data set look like?
- Questions for future steps.

- What information is not self-evident?
- What are different ways you could look at this data?
- How do you plan to slice and dice the data?
- How could you summarize your data to answer key questions?
- What types of plots and tables will help you to illustrate the findings to your questions?
- Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.
- Questions for future steps.

Step 3 - Weeks 11 & 12

You are now on to the final phase of your research paper. While this step does not require you build a model, you are welcome to do so if you feel you have the time. Instead, you need to make a recommendation for the approach you would take and what the remaining steps would be using the information you have learned in this course to take this project from simply being an analysis exercise to proposed implementation of a solution.

- Overall, write a coherent narrative that tells a story with the data as you complete this section.
- Summarize the problem statement you addressed.
- Summarize how you addressed this problem statement (the data used and the methodology employed, including a recommendation for a model that could be implemented).
- Summarize the interesting insights that your analysis provided.
- Summarize the implications to the consumer (target audience) of your analysis.
- Discuss the limitations of your analysis and how you, or someone else, could improve or build on it.

You can use the following template for Step 3:

- A story / narrative that emerged from your data. Follow this structure.

- Introduction.
- The problem statement you addressed.
- How you addressed this problem statement
- Analysis.
- Implications.
- Limitations.
- Concluding Remarks