TidyVerse Trek: Embark on Your Data Odyssey.

Carmel Camilleri

Department of Psychology, York University, Toronto, Canada

GS/PSYC6135 M - Psychology of Data Visualization

Dr. Michael Friendly

April 25, 2024

TidyVerse Trek: Embark on Your Data Odyssey

Within the quantitative methods department at York University, statistical analysis is primarily taught through a statistical software called R (R Core Team, 2024). R is an open-source online software that is very popular in fields with a focus on quantitative methods. R is popular among statisticians and researchers alike because of its extensive library of packages and functions that can be used for data analysis, exploration, and visualization. As well, R offers reproducibility and transparency in the code created. This software has been gaining immense popularity in the last 2 decades as the number of citations of the R Core Team has seen a growth of 87% between 2006 and 2018 (Nordmann et al., 2022). A major advantage of using R is the plethora of options available for data visualization. These visualization options allow users to create beautiful graphs that grab people's attention while being fully transparent about how the data was preprocessed and presented. Although R has many benefits, a huge barrier for users is the level of difficulty associated with syntax-based software like R (Robins et al., 2003).

It is important to distinguish between two popular software styles for data visualization: point and click software and syntax-based software. Point and click software allows users to interact with the program of choice by simply clicking on specific elements, icons or buttons using a mouse or touch screen. Common point and click software like Excel, SPSS and JAMOVI are very user-friendly and easily used. Although users can complete simple tasks using point and click software with a relatively minimal amount of training or expertise, this software lacks the creative and technical freedom that can be found in syntax-based software. Syntax-based software requires users to input specific commands or instructions using a specific syntax or programming language. These commands involve users employing a specific parameter according to a set of rules in a programming language. Some of these languages include R, Python or Javascript. Although this is the recommended approach for both simple and complex analysis due to the researcher's degree of freedom associated with it, there is often a steep learning curve that can be frightening and daunting for new users to pick up and use.

Goals and Objectives for the Workshop

The majority of psychology courses taught in undergraduate and graduate-level statistics courses focus on teaching students how to use the software R for statistical analysis. However, students are often not taught important steps in the analytical process, such as skills in preprocessing data and visualizing their results. As discussed previously, a barrier to this is students' attitudes and self-efficacy toward their coding skills on syntax-based software (Hsu et al., 2009). Despite user's reliance on point-and-click software interfaces, it's often unrecognized that advanced utilization of tools like SAS, SPSS, or Excel demands computational thinking akin to that of syntax-based coding. To fill the gap in the curriculum currently found in the quantitative department at York University, this paper intends to propose a workshop to be run through the quantitative methods department at York University that focuses on preprocessing and visualizing data through the use of R. This Workshop will build fundamental skills in the Tidyverse package to allow users to feel comfortable in preprocessing and visualizing their data. The intended learning outcomes for this course are as follows. (1) Enhance students' confidence and self-efficacy towards approaching and employing preprocessing and visualization strategies in their everyday work, (2) build a collective community of like-minded individuals, and (3) Learn fundamental knowledge in creating visualizations using R.

Why Choose R for Data Visualization?

Data Visualization is vital in the research process but can ultimately be hindered by specific factors often found in point and click software. These limitations are similar to those found when conducting statistical analysis. These limitations include the lack of reproducibility, transparency, and customizability in the user's command. As discussions of a replication crisis in psychology are ongoing, the need for transparency and reproducibility is at an all-time high within the psychological research landscape (Munafò et al., 2017). Also, there is a personal benefit to having reproducibility in users' code that is not discussed as much as the broader repercussions of reproducibility in terms of the replication crisis. The added personal benefit is allowing the user to quickly reproduce a visualization from previous sessions in a split second. This added benefit can save researchers hours of time in reproducing a unique visualization. As the breadth of customization found in this software may appear daunting, an avid online community with solutions and support networks is available to users. For these reasons, global news outlets like the BBC (Journalism, 2019) and *The New York Times* (Bertini & Stefaner, 2015) employ R as their preferred tool for data visualization. R also has an added feature called R Markdown. This document within R places code into specific chunks that are separate from a personal notebook that users can use to write any relevant information (Grayson et al., 2022). This added feature allows users to record their ideas or comments neatly and organized rather than putting a # before text within a document, which can often be confusing and disorganized.

Proposed Workshop Methodology

The quantitative method workshop series (QMWS) will offer this workshop. This series offers a wide range of workshops through the quantitative methods department at York University that is available to anyone of interest. This workshop will be intended to run during the 2024 fall semester. The workshop will be broken down into three distinct sessions. (1) introduction to data preprocessing and exploratory analysis; (2) working with ggplot; (3) practical assignment.

The focus of the workshop will be to foster practical skills. For this reason, each student attending the workshop will have a tailored experience. What is meant by this is that throughout the workshop, students will be building towards the final practical assignment. What is special about this assignment is that students will be asked to bring datasets from their lab or find one relevant to their field of research. This will help users gain a broader understanding of how to manipulate and visualize key variables of interest from their field of study. Throughout the first two sessions, participants will utilize a mtcars dataset. This dataset is built into R and contains information extracted from the 1974 Motor Trend US magazine (*Mtcars Function - RDocumentation*, n.d.).

Peer Instructing Teaching

The teaching approach taken in this workshop will be different than other styles typically taught in the QMWS. This approach will foster a peer instruction approach pioneered by Eric Mazur at Harvard (Mazur & Pearson Education, 2014). Mazurs peer instructed teaching can be understood in 5 distinct steps. (1) The instructor briefly introduces the topic of interest (i.e., R programming for data processing and visualization). (2) The instructor is to give the students a multiple-choice question that probes for misconceptions rather than a simple recall. This could be an understanding of the use of a specific function or package. The ideal question should yield a correct answer rate of 40%-60% (Kober et al., 2015). (3) learners will individually vote on the answer they believe is correct. This helps students form an initial prediction. (4) Students are encouraged to take several minutes to discuss within a small group and practice writing specific code relevant to the question of interest. After this, students come back together and vote individually again. (5) The instructor then takes up the questions; if everyone has the correct answer, the instructor will move on. Any misconceptions are taken up in class-wide discussions to facilitate group learning. I believe that taking this reformed approach to teaching computational thinking and knowledge to a primarily psychological-based community can help radicalize their attitudes and knowledge toward using R for data analysis and visualization.

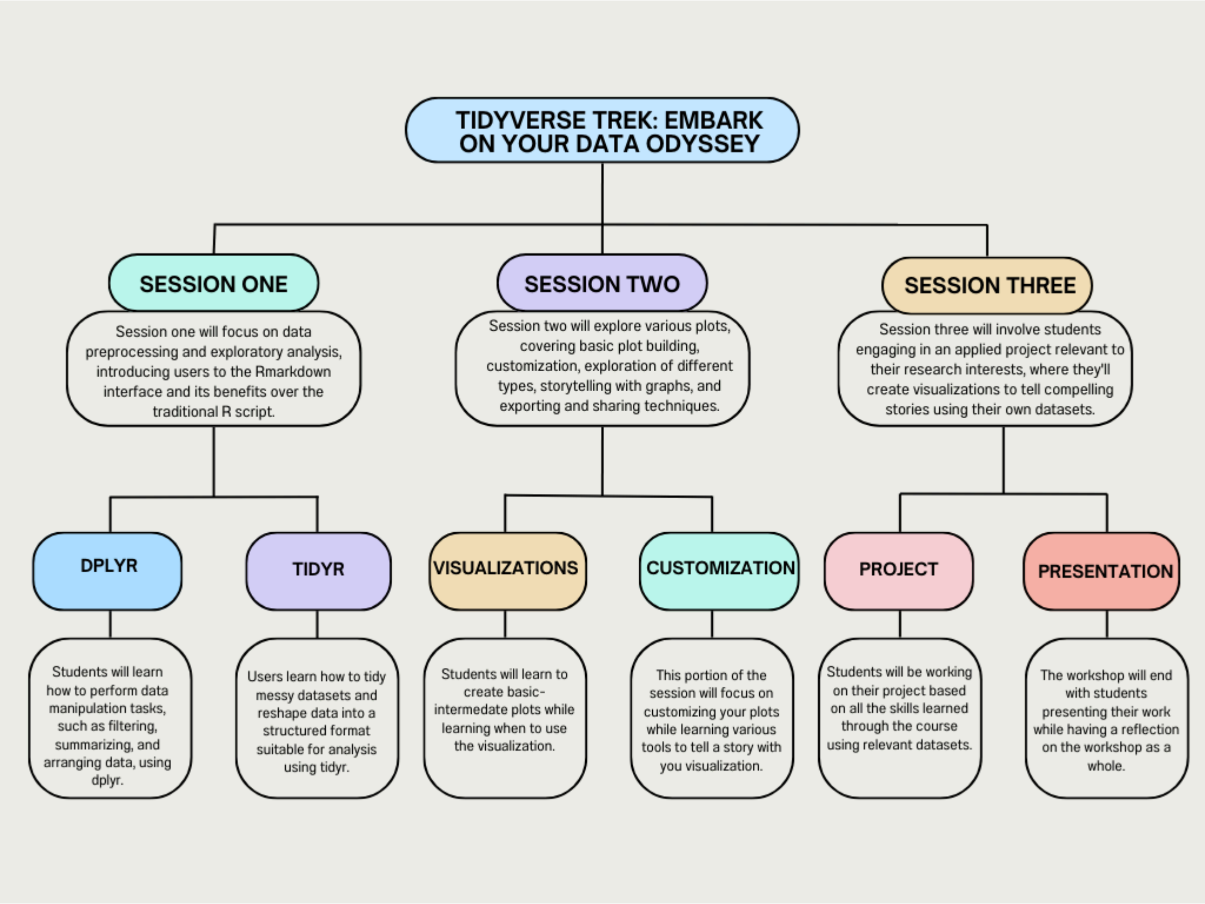
Packages of interest

This workshop will primarily focus on using the `Tidyverse` (Hadley Wickham, 2017). This "composite package" contains additional packages specialized for data science purposes. In particular, the package `ggplot2` will be primarily used for data visualization purposes in this tutorial (Wickham, 2019). Two other main packages will also be employed in the first session of the workshop: `dplyr` and `tidyr`.

Sessions Overview

Figure 1

Workshop diagram overview across three sessions



TIDYVERSE TREK: EMBARK ON YOUR DATA ODYSSEY

SESSION ONE

Session one will focus on data preprocessing and exploratory analysis, introducing users to the Rmarkdown interface and its benefits over the traditional R script.

SESSION TWO

Session two will explore various plots, covering basic plot building, customization, exploration of different types, storytelling with graphs, and exporting and sharing techniques.

SESSION THREE

Session three will involve students engaging in an applied project relevant to their research interests, where they'll create visualizations to tell compelling stories using their own datasets.

DPLYR

Students will learn how to perform data manipulation tasks, such as filtering, summarizing, and arranging data, using dplyr.

TIDYR

Users learn how to tidy messy datasets and reshape data into a structured format suitable for analysis using tidyr.

VISUALIZATIONS

Students will learn to create basic-intermediate plots while learning when to use the visualization.

CUSTOMIZATION

This portion of the session will focus on customizing your plots while learning various tools to tell a story with you visualization.

PROJECT

Students will be working on their project based on all the skills learned through the course using relevant datasets.

PRESENTATION

The workshop will end with students presenting their work while having a reflection on the workshop as a whole.

Session one

Session one will focus on data preprocessing and exploratory analysis. Users will be taught the Rmarkdown interface and the many benefits of using this in a learning environment over the normal R script. After spending time on this topic, the remainder of the session will focus on diving into two distinct packages within the composite `tidyverse`: `dplyr` and `tidyr`. These packages play a pivotal role in data manipulation and offer a comprehensive toolkit for transforming messy data into tidy, ready-to-analyze datasets. This session will first focus on data manipulation and filtering. Users will learn unique functions like `filter()and select()` to subset and manipulate variables of interest. Next, students will take a deeper dive to learn data summary for relevant groups of interest. Functions like `summarize()` and `group_by()` will be taught. Next, variable manipulation and data sorting will be taught to students. Relevant functions like `mutate()` and `arrange()` will be discussed. From here, students will work with separate datasets with unique identifiers of interest. A focus on data joining will now be explored, and students will learn how to employ vital functions such as `left_join(), right_join(),` and `inner_join()`. Data separating and uniting will be the session's 2nd last topic of interest. Here, users will learn unique functions like `separate()` and `unite()` to dive deeper into data manipulation. Lastly, users will learn the basics of missing values, and functions like `drop_na()` will be used.

Session Two

Session two will primarily focus on learning how to utilize R to make basic to intermediate visualizations. This section will be broken down into 5 distinct sections: (1) building basic plots, (2) customizing plot appearance, (3) exploring different plot types, (4) telling a story with your graph, and (5) exporting and sharing your visualizations. For the first section, students will learn the basic structure of the ggplot2 syntax and how to map variables using aesthetics. This will be built upon to demonstrate how to add geometric (geom) objects to represent data. Here, users will learn how to plot simple visualizations such as a scatter plot, line chart or a bar graph. The second section will contain valuable information for customizing plot appearance. This section will highlight how to modify plots for consistency and clarity (i.e., titles, axis labels, and legends). Also, aesthetic customization options such as colour, shape, and size of various elements will be explored further. The third section will contain information regarding constructing more advanced plot types such as histograms, density plots, boxplots, violin plots, and jitter plots. In this section, users will take the information learned in the first session to understand how to summarize data and transform that information into elegant visualizations. The next section will teach users how to tell a story using their visualization. This will include selecting the appropriate visualization for a particular type of data to create a narrative-driven visualization. The last section will teach users how to correctly save their plots in various formats (e.g., PNG, JPEG). This section will also instruct users on best practices for presenting and sharing their visualizations.

Session three

The last session will primarily consist of an applied project relative to the research of the students attending. This section is novel, as users are tasked with working on a project relevant to their area of interest. In this project, users will have the opportunity to subset, create or modify their dataset to tell an intriguing visualization-driven story to the class. This will foster a community among students as each person will have the opportunity to learn about each other's research through various visualizations. Each user will have to create a visual dashboard to convey a particular story. There is no marked grade, only participation.

The last part of this session will be an open discussion to see what users thought was the most beneficial in the course and what they would like to learn more about in the future.

Conclusion

The proposed workshop represents an innovative teaching opportunity for students who are keen on attending the QMWS. First, the inclusion of this workshop will focus on developing hard skills (i.e., data exploration and visualization using R) not taught in the current curriculum in the QM department. Secondly, this workshop will employ novel teaching methods to advance the comprehension and overall attitudes of students who are not familiar with or confident in their R skills. Lastly, this course will primarily foster users' visualization skills using R in their relevant field of study. Hopefully, this will encourage students from a breadth of research backgrounds to join the workshop to improve their exploratory data analysis skills and visualization storytelling abilities.

References

56 | Amanda Cox on Working With R, NYT Projects, Favorite Data. (2015, June 25). Data Stories. https://datastori.es/ds-56-amanda-cox-nyt/

Bolk, J., Simatou, E., Söderling, J., Thorell, L. B., Persson, M., & Sundelin, H. (2022). Association of Perinatal and Childhood Ischemic Stroke With Attention-Deficit/Hyperactivity Disorder. *JAMA Network Open*, *5*(4), e228884. https://doi.org/10.1001/jamanetworkopen.2022.8884

Bertini E., Stefaner M. (2015). Amanda Cox on working with r, NYT Projects, Favorite Data [Podcast]. *Data Stories*. https://datastori.es/ds-56-amanda-cox-nyt/

Grayson, K. L., Hilliker, A. K., & Wares, J. R. (2022). R Markdown as a dynamic interface for teaching: Modules from math and biology classrooms. *Mathematical Biosciences*, *349*, 108844. https://doi.org/10.1016/j.mbs.2022.108844

Hadley Wickham. (2017). Easily Install and Load the "Tidyverse" [R package tidyverse version 1.2.1]. *R-Project.org*. https://cran.r-project.org/package=tidyverse

Hsu, M. K., Wang, S. W., & Chiu, K. K. (2009). Computer attitude, statistics anxiety and self-efficacy on statistical software adoption behavior: An empirical study of online MBA learners. *Computers in Human Behavior*, *25*(2), 412–420. https://doi.org/10.1016/j.chb.2008.10.003

Journalism, B. V. and D. (2019, February 1). *How the BBC Visual and Data Journalism team works with graphics in R*. BBC Visual and Data Journalism. https://medium.com/bbc-visual-and-data-journalism/how-the-bbc-visual-and-data-journalism-team-works-with-graphics-in-r-ed0b35693535

Kober, N., National Research Council (U.S.). Board On Science Education, & National Research Council (U.S.). Division Of Behavioral And Social Sciences And

Education. (2015). *Reaching students : what research says about effective instruction in undergraduate science and engineering*. National Academies Press.

Mazur, E., & Pearson Education. (2014). *Peer instruction : a user's manual*. Pearson Education.

mtcars function - RDocumentation. (n.d.). Www.rdocumentation.org. https://www.rdocumentation.org/packages/datasets/versions/3.6.2/topics/mtcars

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1). https://doi.org/10.1038/s41562-016-0021

Nordmann, E., McAleer, P., Toivo, W., Paterson, H., & DeBruine, L. M. (2022). Data Visualization Using R for Researchers Who Do Not Use R. *Advances in Methods and Practices in Psychological Science*, *5*(2), 251524592210746. https://doi.org/10.1177/25152459221074654

RStudio Team. (2024). *RStudio: Integrated development environment for r*. RStudio, PBC. http://www.rstudio.com/

Robins, A., Rountree, J., & Rountree, N. (2003). Learning and Teaching Programming: A Review and Discussion. *Computer Science Education*, *13*(2), 137–172. https://doi.org/10.1076/csed.13.2.137.14200

Wickham, H. (2019). *Create Elegant Data Visualisations Using the Grammar of Graphics*. Tidyverse.org. https://ggplot2.tidyverse.org