Data Analysis Onboarding📊

Using R

R is a programming language commonly used for statistical analysis, computing, and graphics. While there are many ways to use R, using the tidyverse packages is highly recommended as it has been designed for data science.

Learn more about tidyverse: here, and here.

The creator of the tidyverse packages, Hadley Wickham, also has a great tutorial for R that goes over importing data, transforming it, visualizing it, modelling it, and more. I highly recommend that those unfamiliar with R take a look: here.

In particular, I highly recommend reading the sections on Data Import, Tidy Data, Pipes, Model Basics, and Graphics for Communication.

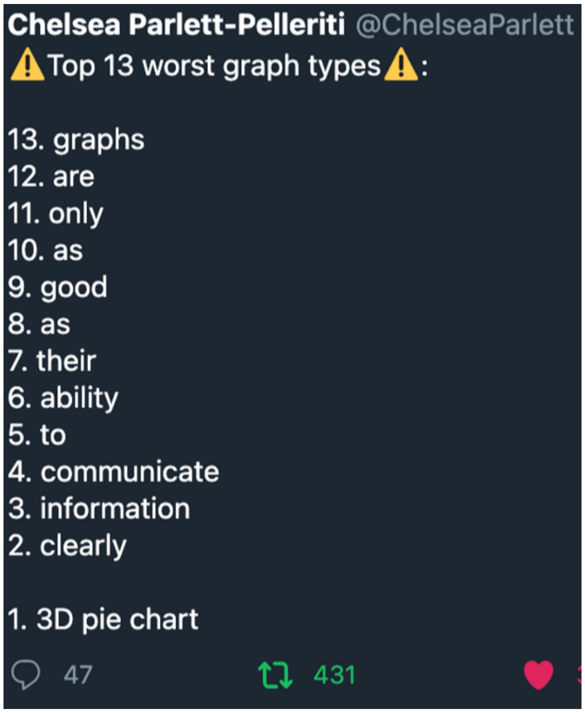
ggplot2

ggplot2 is one of the packages within the tidyverse set used for data visualization, based on the Grammar of Graphics. It is incredibly flexible and intuitive to use, through its semantic scales and layers.

A great introductory tutorial for ggplot2 can be found here. Notably, types of figures you should be familiar with include bar charts[1], histograms, and boxplots. You should also be familiar with adding error bars, and adding text to plots.

Data Visualization Theory

Understanding the theory and principles behind data visualization are also very important to any data analyst, to ensure you are communicating your data efficiently. While there is no "right" or "wrong" graphs to present your data, there are certain types that are more efficient depending on the information you wish to convey.



Chelsea Parlett-Pelleriti @ChelseaParlett

⚠Top 13 worst graph types⚠:

13. graphs
12. are
11. only
10. as
9. good
8. as
7. their
6. ability
5. to
4. communicate
3. information
2. clearly

1. 3D pie chart

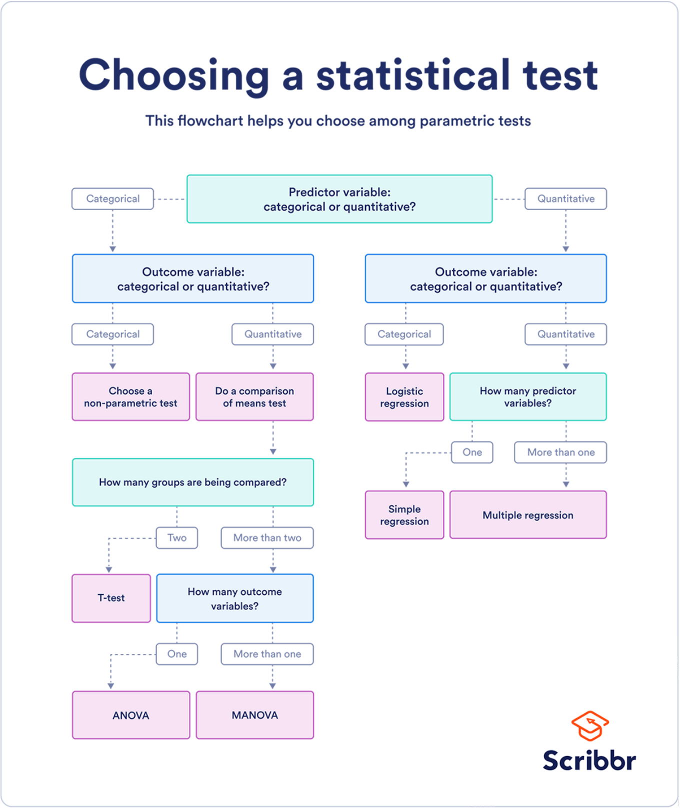
47 431

Please watch this video on Introductory Visualization Analysis & Design and this video on Marks and Channels. For those looking to learn more, this textbook, Visualization Analysis and Design by Tamara Muzner, is also a great resource for those looking to learn more about visualization theory.

[1] Recall it is lab SOP to include SEM bars with bar charts.

Statistical Analysis

A crucial aspect of data analysis is selecting the appropriate test to conduct your analysis. There are a wide selection, and oftentimes there is more than one appropriate test. It is important to consider the nature of your data, your outcome measure, and the story you wish to tell.


Choosing a statistical test
This flowchart helps you choose among parametric tests
Predictor variable: categorical or quantitative?
Categorical
Quantitative
Outcome variable: categorical or quantitative?
Outcome variable: categorical or quantitative?
Categorical
Quantitative
Categorical
Quantitative
Choose a non-parametric test
Do a comparison of means test
Logistic regression
How many predictor variables?
One
More than one
Simple regression
Multiple regression
How many groups are being compared?
Two
More than two
T-test
How many outcome variables?
One
More than one
ANOVA
MANOVA
Scribbr

https://www.scribbr.com/statistics/statistical-tests/

Some commonly used tests in this study are T-tests, ANOVA, logistic/linear regression, and their non-parametric (or non-normal)[2] counterparts. It is worthwhile to become familiar with these and how to implement them in R.

[2] Non parametric tests are also common in this study, as often your response variable can be a likert scale, or non-normal

In this study, oftentimes you are measuring students over multiple weeks/assignments, and thus a repeated measures structure is very common. Repeated measures require special considerations because of the dependency between observations within a particular student. You may consider repeated measures ANOVA, mixed effects models, Friedman tests, etc.

Considerations to have with repeated measures in R

Oftentimes when working with repeated measures data in R, you may have to format your data into long data (read more about long vs wide data here). For instance, instead of having each row in your dataset be a student, it may be one mark on an assignment. A tutorial on converting between long and wide data in R can be found here.

Ethical Statistical Analysis

There are many misconceptions about data analysis, which can sometimes lead to unintentionally (or sometimes intentional) misleading results and conclusions. As a data analyst, you act as the spokesperson for the data, and it is important that you practice ethical statistical analysis. Liza Bolton, a former UofT professor, has a great reading on doing ethical data analysis which you can read here. Please also read this article, on reproducible and transparent data analysis.

—STOP HERE—

Data Documentation

Some resources to better understand the documentation and the structure of the datasets:

Analysis Documentation

OnTrack Fall 2021 - Data Documentation

Replication

Some past analysis to replicate: Winter 2022 Analysis to date

- Code can be found here

For replication, please ask Naaz for access to the data on Gitlab. Then do the following exercises:

1) Create a graph that represents the proportion of students who self-explained by voice in week 9 for each treatment condition.

2a) Find the number of people who rated each score -3 (strong disagree), -2, -1, 0, +1, +2 +3 (strongly agree), for each of these **pre-survey** Likert items, and represent it using this template:

- Q831: When communicating with my friends and family, I prefer calling them on the phone and sending a voice note rather than texting them
- Q832: When communicating with instructors or TAs during lecture or office hours, I prefer using voice (e.g., unmuting and talking) rather than text (e.g., using Zoom chat).

2b) Find the number of people who rated each score -3 (strong disagree), -2, -1, 0, +1, +2 +3 (strongly agree) for each of these **post-survey** Likert items, and represent using this template:

- Q831: When communicating with my friends and family, I prefer calling them on the phone and sending a voice note rather than texting them
- Q832: When communicating with instructors or TAs during lecture or office hours, I prefer using voice (e.g., unmuting and talking) rather than text (e.g., using Zoom chat).

3) In the post-survey, we asked students for their preferences about their preferred reflection medium on a scale from -5 to +5, which are represented by survey items: “I would have preferred to only provide *text* reflections” (Q843_1), “I would have preferred to only provide *voice* reflections” (Q843_3) and “I would have preferred always having an option between providing voice or text reflections” (Q843_2). Create a box plot that shows the ratings of people who preferred each medium type, based on their original treatment condition.

4) Recall the self-explanations are assigned from week 4 to 9. Report the completion rates of the self-explanations for each of the three conditions (voice, text, and choice of explanation medium) per week:

Explanation	Week 4		Week 5		Week 6		Week 7		Week 8		Week 9
	E1	E2	E1	E2	E1	E2	E1	E2	E1	E2	E1
Overall completion rate (N=247)	83.47%	80.25%	83.87%	82.67%	81.86%	82.67%	87.10%	87.91%	74.19%	73.78%	83.87%
Text (N=81 of 247)	28.63%	27.42%	29.44%	29.44%	29.44%	29.44%	30.65%	30.65%	26.61%	26.61%	0%
Voice (N=83 of 247)	27.02%	25.81%	26.61%	26.21%	25.40%	26.21%	27.82%	28.63%	22.58%	22.98%	0%
Choice of Medium (N=83 of 247)	27.82%	27.02%	27.82%	27.02%	27.02%	27.02%	28.63%	28.63%	25.00%	24.19%	83.87%