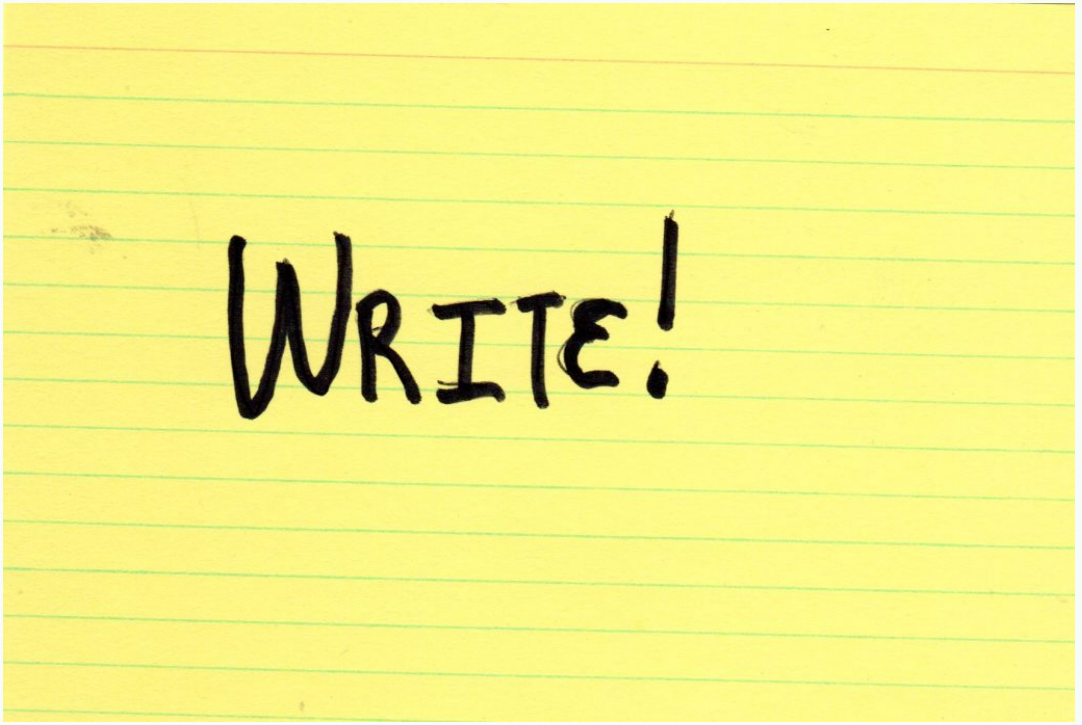
Introduction to Statistics

Stat 120

March 26 2023

Your introduction

- *Your name?*
- *What gender pronouns do you use?*
- *Favorite Scientist/Person?*
- *Recent fun memories?*



Please fill in!

What will you learn in this course?

- Analyzing data by doing exploratory data analysis
- Estimate some parameter of interest from the population
- Infer the population characteristics based in your estimation
- Quantify the uncertainty in the estimation

What will a typical day/week look like?

Before Class:

- Some reading/video to introduce some topics
- Will be updated in the weekly planner

During Class:

- Mini-lectures
- Hands-on group activities and quizzes

💫Wad up

💫Wad up

💫Not much

💫Wad up

💫Not much

💫Not much again

💫Wad up

💫Not much

💫Not much again

💫Not much when

💫Wad up

💫Not much

💫Not much again

💫Not much when

💫Not much then

💫Wad up

💫Not much

💫Not much again

💫Not much when

💫Not much then

💫Not much then again

Statistics is distinct from mathematics

Statistics is the study of data and the uncertainties surrounding them. We will take a more conceptual route to statistics in this course.

Statistical Jargons

- Knowing basic statistical jargon is essential to understand the results of a statistical analysis.
- Statistical jargon can be used to help make decisions based on data, such as whether or not to reject or accept a hypothesis or to determine if a correlation exists between two variables.
- Understanding statistical terms can help to effectively communicate the findings of a statistical analysis.

What and Why of Statistics?

Science of collecting, describing, analyzing and making decisions based on data

- ***Sampling***
- ***Exploratory Data Analysis***
- ***Inference***

Allows us to make informed decisions in the face of uncertainty and let's us take an unbiased and evidence-based viewpoint

Data: Cases and Variables

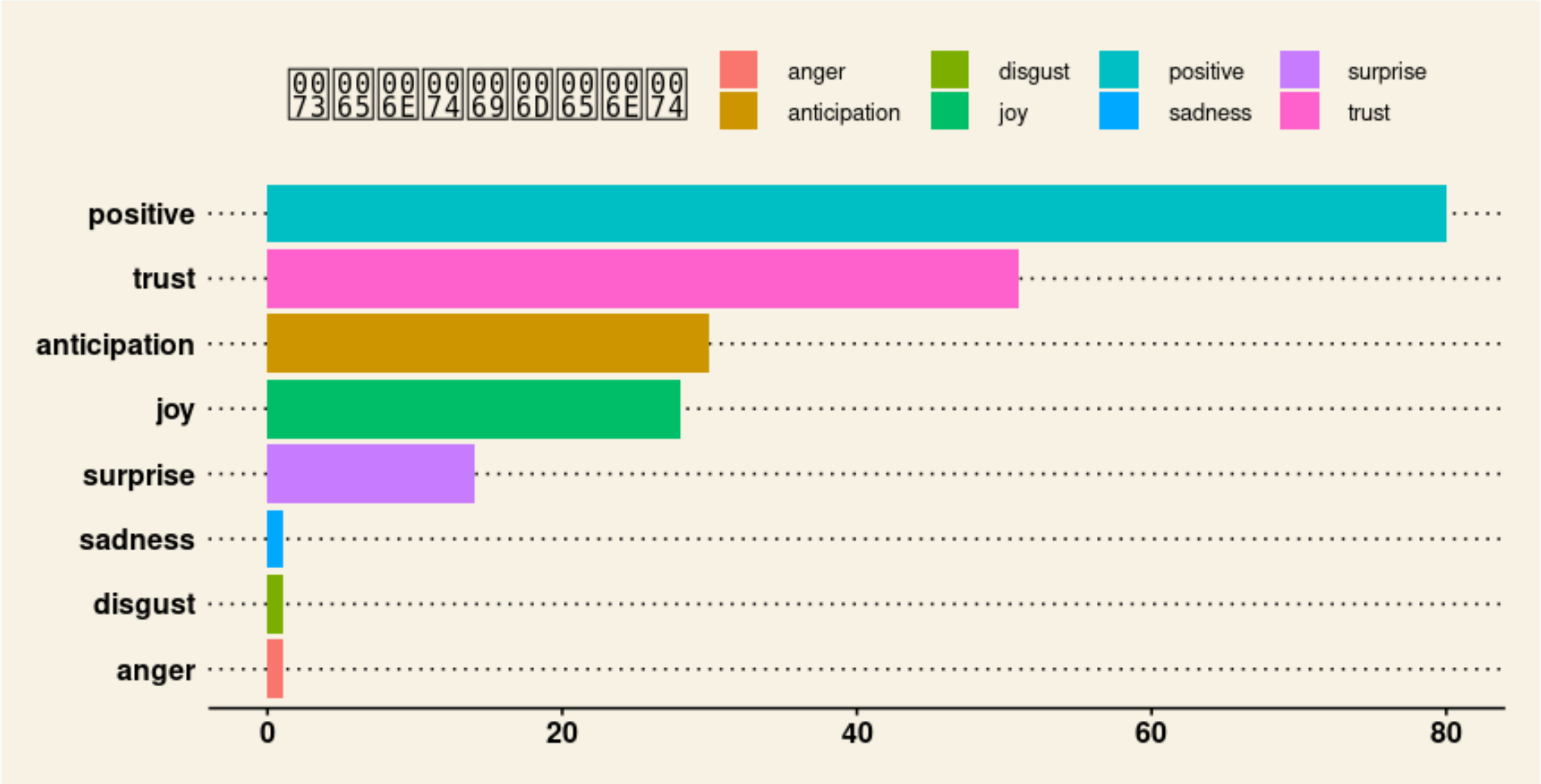
Data are a set of measurements taken on a set of individual units

- *These are **cases** or units*

Data is stored and presented in a dataset that comprises of variables measured on cases

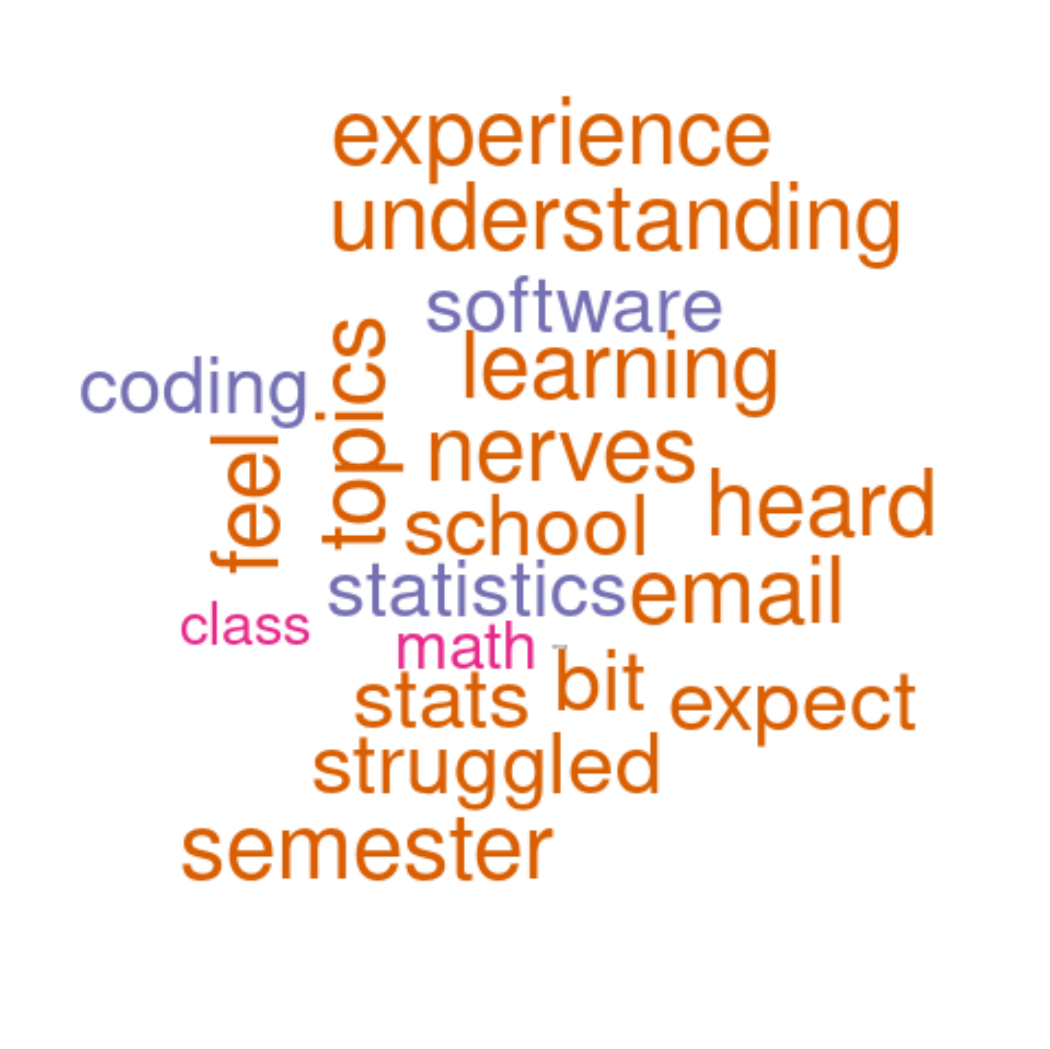
- *A **variable** is any characteristic that is recorded for each case*

General sentiment of the class



Survey Responses: What do you hope to gain from taking this course?

Survey Responses: Is there anything about this course that you are nervous about?



Software Component: We will gradually learn these things!

- Statistical computing software called R
- RStudio gives nice user-friendly interface to R
- RMarkdown is platform in Rstudio to write your codes and results

`EducationLiteracy` dataset from Lock5

Country	Code	Education	Literacy
Afghanistan	AFG	4.23	43.0
Albania	ALB	3.95	98.1
Algeria	DZA	NA	81.4
Andorra	AND	3.26	NA
Antigua and Barbuda	ATG	NA	99.0
Argentina	ARG	5.78	99.2
Armenia	ARM	2.81	99.7

Each row = case & Each column = variable

Click here for Lock5 datasets for the second edition

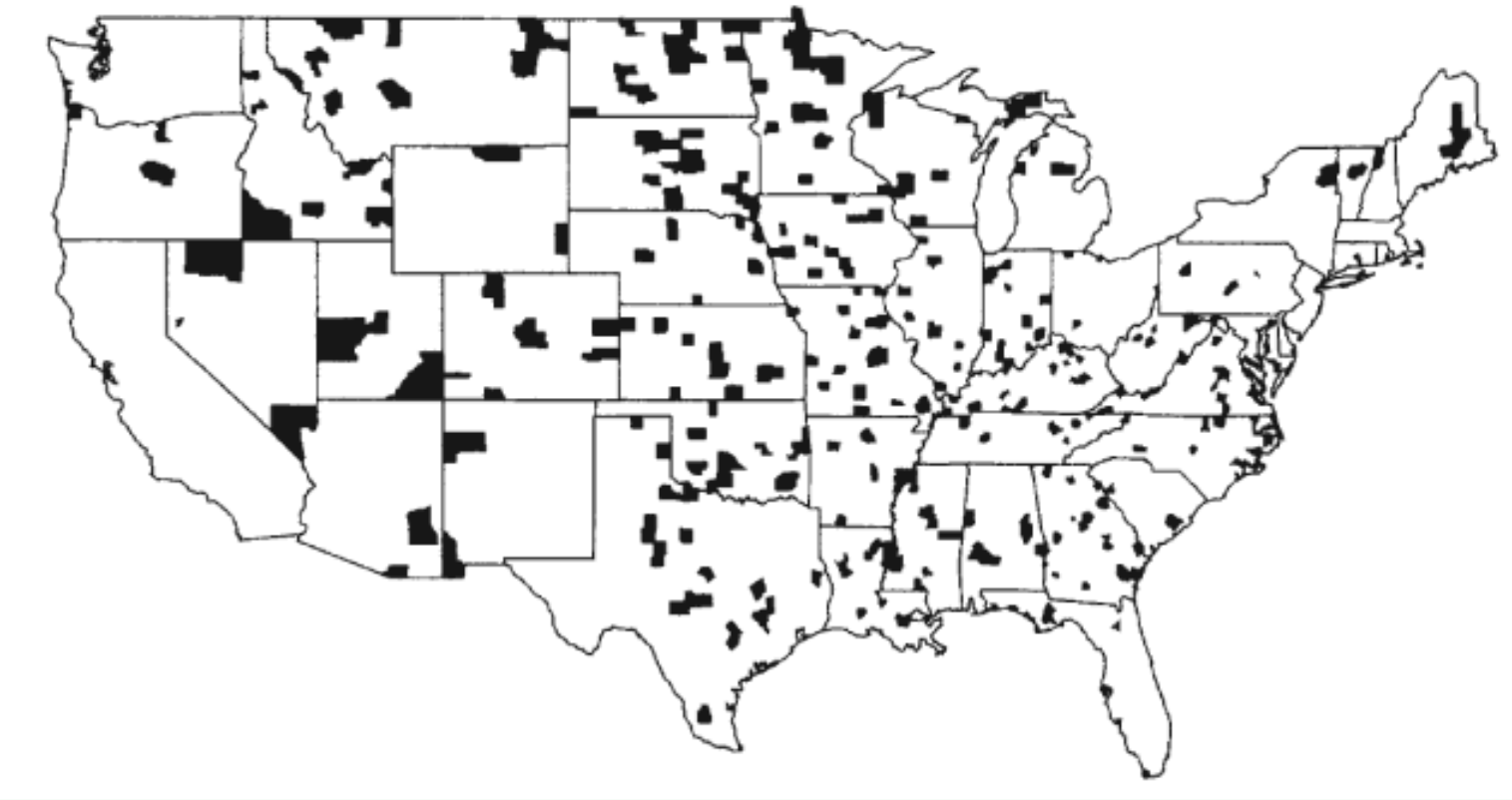
Categorical Versus Quantitative

Variables are classified as either categorical or quantitative:

- ***A categorical variable divides the cases into groups. e.g. gender, country, state etc.***
- ***A quantitative variable measures a numerical quantity for each case, e.g. age, height, sleep hours, blood pressure etc***

Kidney cancer

Counties with the highest kidney cancer rates



Source: Gelman et. al. Bayesian Data Anaylsis, CRC Press, 2004

Kidney cancer

If the cases in the kidney cancer dataset are people, then the measured variable is categorical

- We categorize each person as either having kidney cancer or not which is categorical.

Kidney cancer

If the cases in the kidney cancer dataset are counties, then the measured variable is quantitative

- Data collected at the county level is aggregated across all people living in the county. We then get rates of cancer which are numbers (quantitative).

Variable manipulations

Can use numbers to code categories of categorical variable

- *e.g Gender (1 for male and 2 for female)*

Can convert quantitative variable into categorical groups

- *e.g. Income (0-50000 as Low, 50000+ as High)*

Categorical variables are sometimes collapsed into fewer levels

- *e.g. no HS degree, HS degree, some college, 2 year degree*

Explanatory and Response Variable

When one variable helps us understand or predict values of another variable, we call the former the explanatory variable and the latter the response variable

Does meditation help reduce stress?

- *explanatory variable: meditation*
- *response variable: stress level*

Does sugar consumption increase hyperactivity?

- *explanatory variable: sugar consumption*
- *response variable: hyperactive behavior*

BuzzR

RStudio anatomy

https://buzzrbeeline.blog/

Emma Rand

Script file

Write code here

To run code put your cursor on the line and click the run button

Edit to correct errors

⇒ record of commands that worked

Save scripts with the **.R** extension

⇒ syntax will be highlighted

⇒ good practice

`<-` is the assignment operator

⇒ puts what is on the right in to the object on the left

⇒ Assign results if you want to use them again

Console

When you click run, code is sent to the console and executed

`>` is the prompt

⇒ do not type it

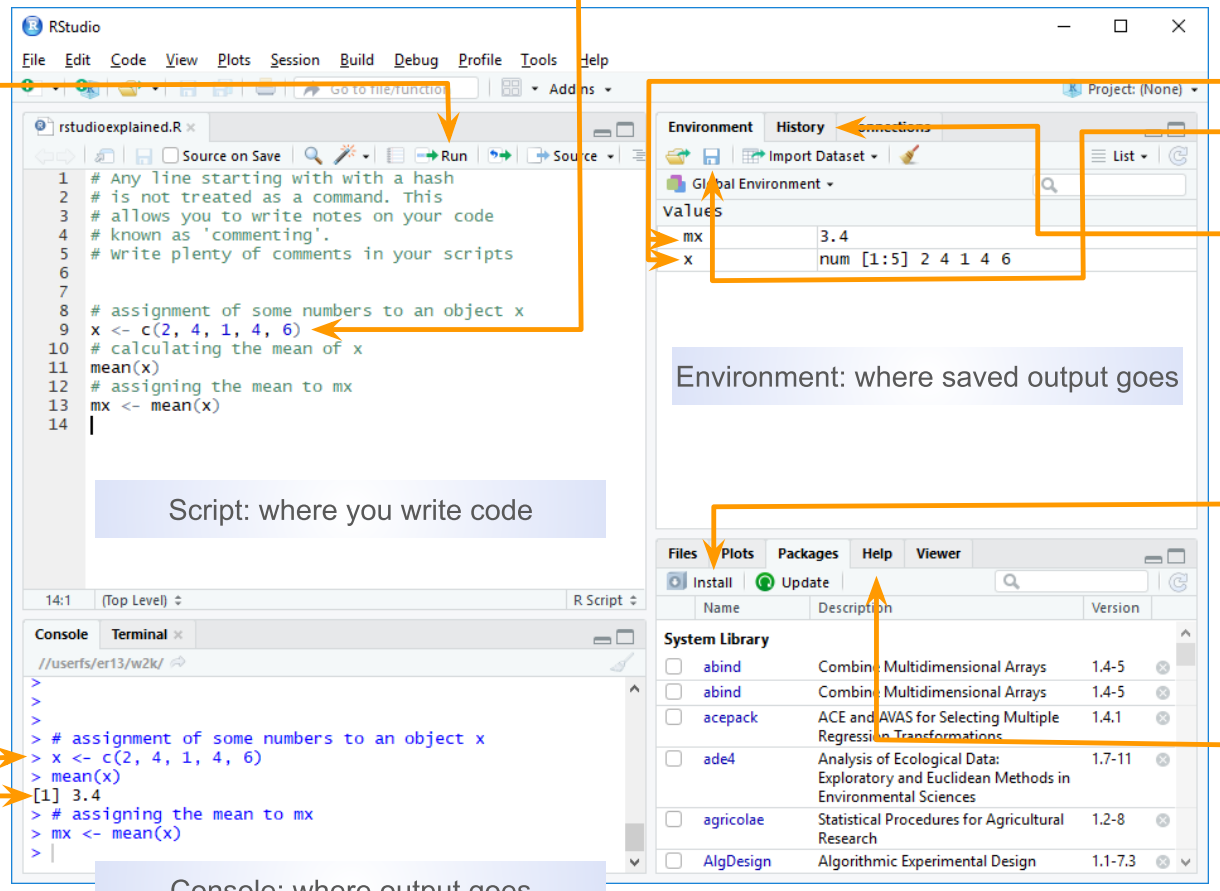
⇒ appears when R is ready for next command

Command output goes here by default

⇒ output is in a different colour

⇒ `[1]` indicates 3.4 is the first element of the output

⇒ many commands will not have output, the prompt just reappears



Environment

Name objects by assignment to use them again

All the objects you created in your session

Saving the environment saves all the objects, but not the code with a **.RData** extension

History

A history of every command you sent to the console, mistakes included.

File can be saved but usually you just need the script

Packages

Many functions come with R

A huge amount of extra functionality is available in packages

Packages can be installed by clicking the Install button

Help

Access to manual pages for all installed packages

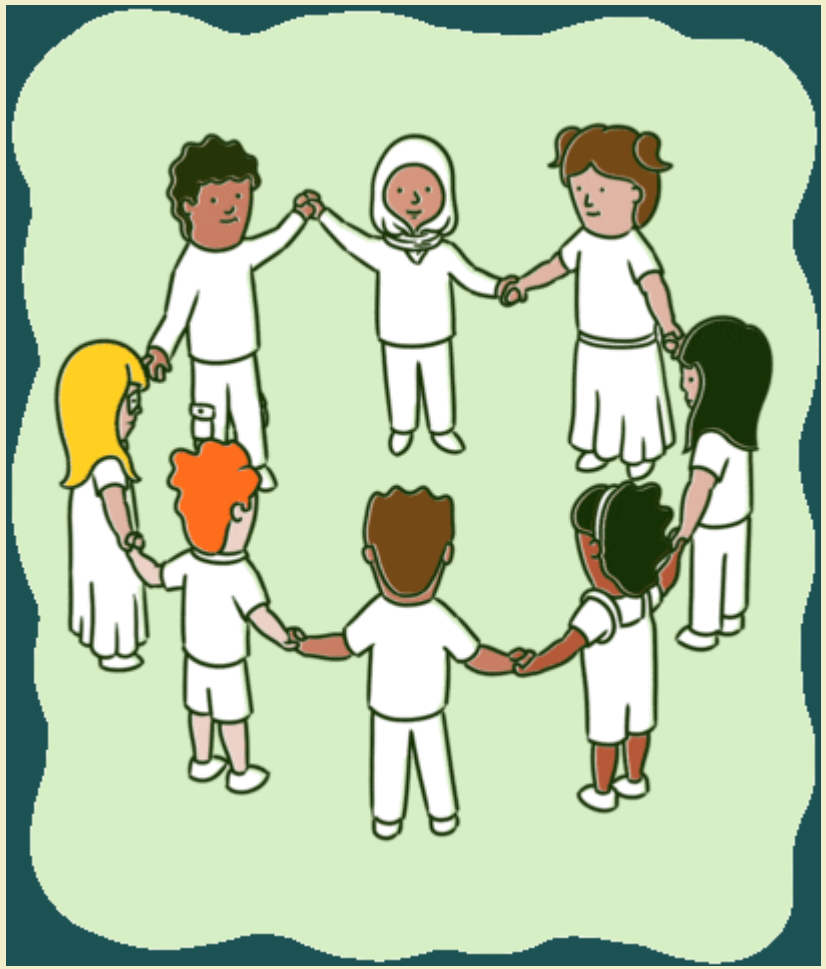
Plots

Figure output appears here

Click here for source

Group Activity 1

10:00



- Make a course folder called 'stat120' either on your Maize account or on your local computer
- Please download the in-class activity file for Day 1 from moodle and go to class activity module
- We will go over this .Rmd file together.
- Once you are done, knit to pdf or word and submit.