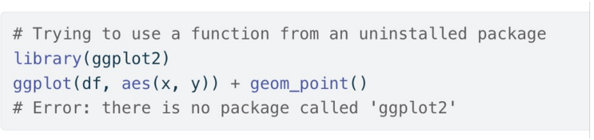
Common challenges and resources

How do you start exploring a new dataset?

- How should I load the data file?
- What type of data are we looking at?
    - What is the size of the dataset?
    - What are the variables in the dataset?
    - Is the data complete? Is incompleteness on purpose?
- What are the question(s) I’m trying to address?

How should I load the data file?

- Look at the file extension → *Do you recognize it?*
  - No? Google “what is a **txt** file extension” (others: csv, tsv, Rda, Rds)
- Is a package required to look at this data?
  - Large data files: **data.table** package → **fread**()
  - Excel files: **readxl** package → **read_excel**()
  - Google → “What package can be used to read a **txt** file in R”
- Is the package installed?
  - install.packages(“package_name”)
  - Github packages may require the **devtools** package for installation
  - Bioconductor packages provide the code for package installation
  - Google → “How do I install package **package_name** in R”



```r
# Trying to use a function from an uninstalled package
library(ggplot2)
ggplot(df, aes(x, y)) + geom_point()
# Error: there is no package called 'ggplot2'
```

What does the data look like?

- Does it need to be “cleaned”?
  - Should outliers or bad samples be removed (filtered out)?
    - Idea: Use PCA to identify samples that don’t look like the rest of your cohort
  - Do I care about all of the data or just some of it?
  - Did the experiment go well?
- Identifying missing data or sparse data

Missing data (vs. sparse data)

Why might data be missing?

- The file you read doesn’t contain values for a given column in a subset of rows
- You performed some computation and the value has become to small or large to represent in memory.
- You tried to add a column incorrectly.
- Technical reasons

Sparse data is data that contains a significant amount of 0.

- Efficient representation

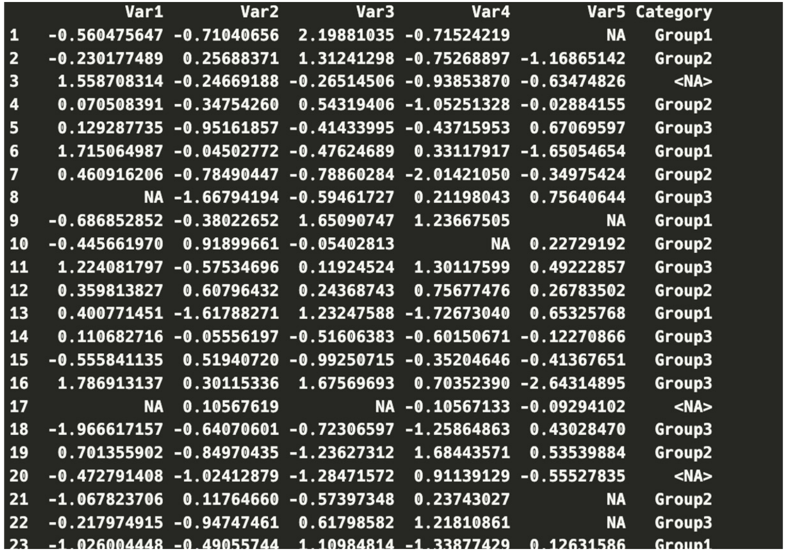
Missing Data Representation

How is this represented in R?

- NA (Not available/applicable)
  - as.character(NA)
- NaN (Not a Number)
  - Undefined or unrepresentable result from some computation
- NULL (No value)
  - Value never existed
- INF/-INF

How can we overcome this?

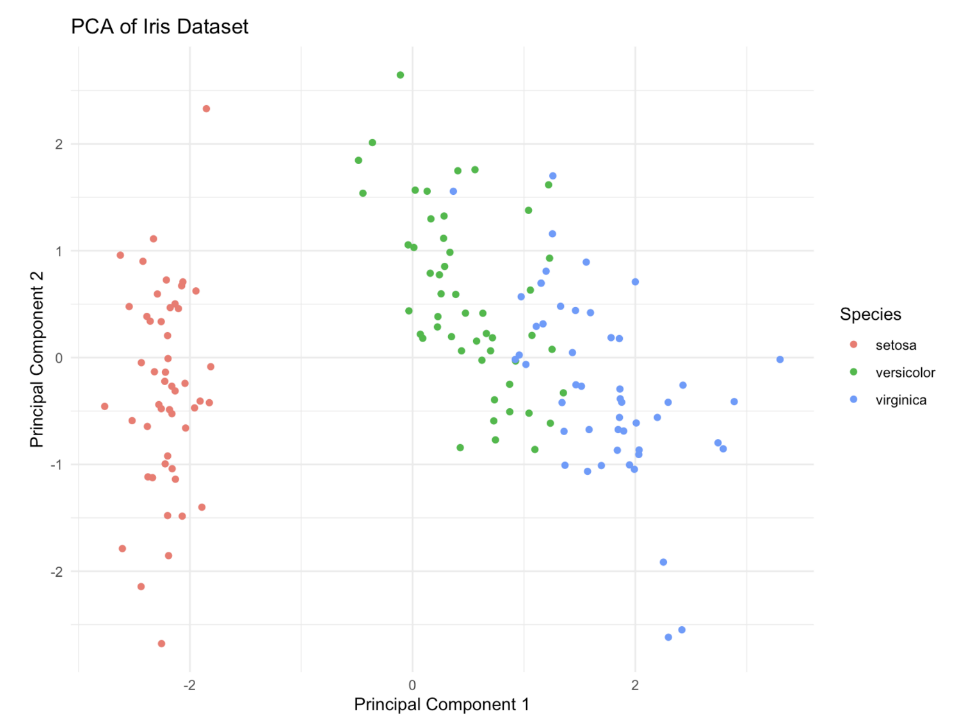
- na.rm parameter (statistics functions)
- na.value parameter (ggplot2)
- Filtering NA values



```
           Var1        Var2        Var3        Var4        Var5 Category
1  -0.560475647 -0.71040656  2.19881035 -0.71524219          NA   Group1
2  -0.230177489  0.25688371  1.31241298 -0.75268897 -1.16865142   Group2
3   1.558708314 -0.24669188 -0.26514506 -0.93853870 -0.63474826     <NA>
4   0.070508391 -0.34754260  0.54319406 -1.05251328 -0.02884155   Group2
5   0.129287735 -0.95161857 -0.41433995 -0.43715953  0.67069597   Group3
6   1.715064987 -0.04502772 -0.47624689  0.33117917 -1.65054654   Group1
7   0.460916206 -0.78490447 -0.78860284 -2.01421050 -0.34975424   Group2
8            NA -1.66794194 -0.59461727  0.21198043  0.75640644   Group3
9  -0.686852852 -0.38022652  1.65090747  1.23667505          NA   Group1
10 -0.445661970  0.91899661 -0.05402813          NA  0.22729192   Group2
11  1.224081797 -0.57534696  0.11924524  1.30117599  0.49222857   Group3
12  0.359813827  0.60796432  0.24368743  0.75677476  0.26783502   Group2
13  0.400771451 -1.61788271  1.23247588 -1.72673040  0.65325768   Group1
14  0.110682716 -0.05556197 -0.51606383 -0.60150671 -0.12270866   Group3
15 -0.555841135  0.51940720 -0.99250715 -0.35204646 -0.41367651   Group3
16  1.786913137  0.30115336  1.67569693  0.70352390 -2.64314895   Group3
17           NA  0.10567619          NA -0.10567133 -0.09294102     <NA>
18 -1.966617157 -0.64070601 -0.72306597 -1.25864863  0.43028470   Group3
19  0.701355902 -0.84970435 -1.23627312  1.68443571  0.53539884   Group2
20 -0.472791408 -1.02412879 -1.28471572  0.91139129 -0.55527835     <NA>
21 -1.067823706  0.11764660 -0.57397348  0.23743027          NA   Group2
22 -0.217974915 -0.94747461  0.61798582  1.21810861          NA   Group3
23 -1.026004448 -0.49055744  1.10984814 -1.33877429  0.12631586   Group1
```

How can ChatGPT *help* (but it won't do everything for you)

Live demo: How do I figure out how to make these two clusters?



Live demo: How do I replicate the analysis performed in a figure?

Create an analysis strategy by working backwards.

Deconstructing someone else’s plot

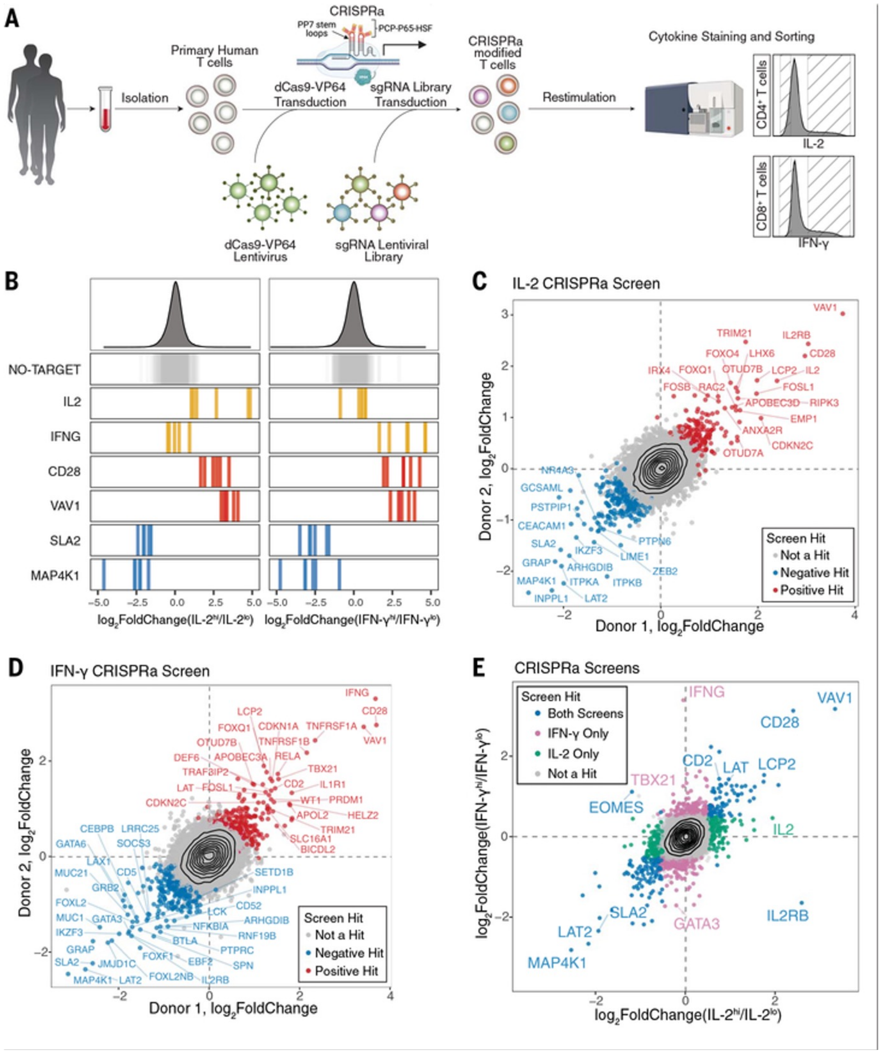


Figure 1
Schmidt, et al. *Science* 2022

Hands-on: Ready to take on additional datasets

- These slides are available on the course website under **Common challenges and additional resources**
- Coding
    - Work through the blocks of code under the **Course: Common challenges and additional resources** page
        - Review the details on how the code works in the Lecture slides for assistance
        - Put a post-it on your laptop if you get stuck, indicating for a TA to come up to you
        - Work through the blocks of code on this page, practicing in both your Rscript and the terminal
    - Taking the next step
        - There are a list of **Additional exercises** at the bottom of the page for you to try on your own

Goal: Are you ready to apply your foundational skills to your own data?