

Introduction to Data Analysis

Lecture 01

Agenda

- Course content
- Motivation for R in data analysis

Course description

- This is a course for students and researchers who have data or starting to gather data in their research work.
- This course deals on how to handle, manage, analyze, and visualize data in health and biomedical research.

Course objectives

1. To answer research questions using data
2. To learn basic concepts in data wrangling, exploratory data analysis, statistical analysis, reproducible research, and data communication for publication
3. To write codes involving real life datasets using the R language

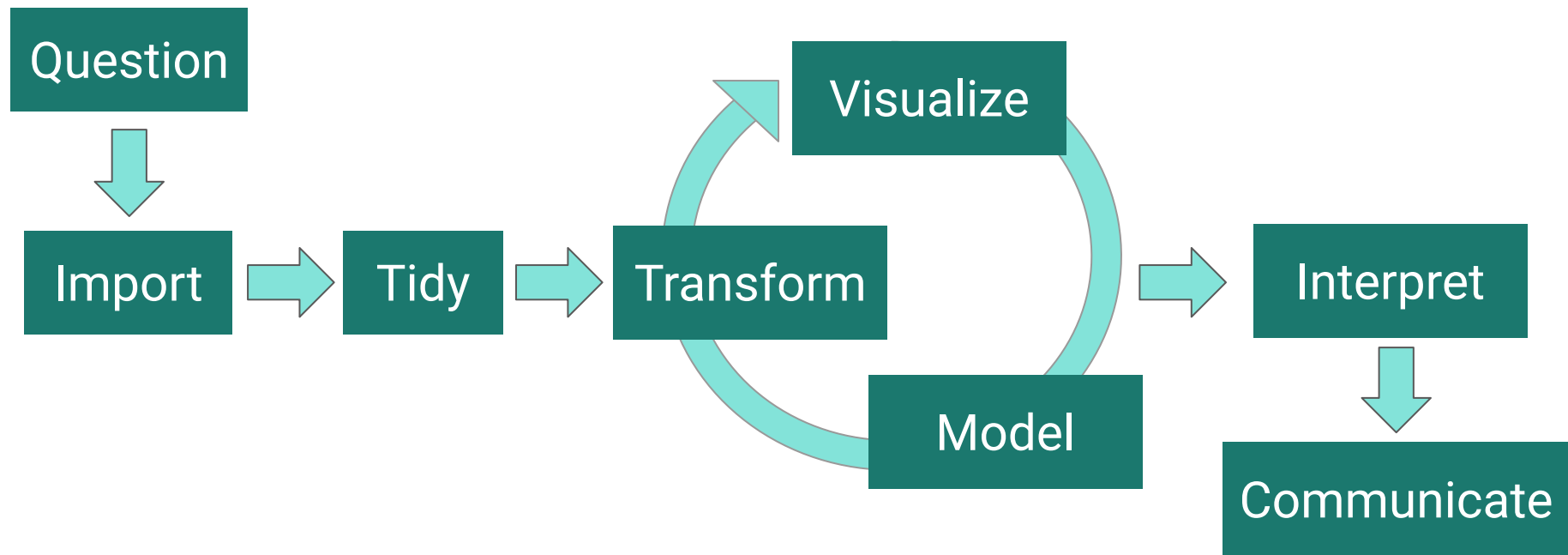
Course format

- Classes will be conducted online via Google Meet
- Lectures with hands-on exercises
- Mondays, 10:00~11:00 AM

Course schedule

Week	Date	Topic
1	07 Dec 2020	Introduction to Data Analysis
2	14 Dec 2020	From Variables to Data Frames
3	21 Dec 2020	Control Flow
4	18 Jan 2021	Importing Data
5	25 Jan 2021	Cleaning Data
6	01 Feb 2021	Tidying Data
7	08 Feb 2021	Transforming Data
8	15 Feb 2021	Merging Data
9	22 Feb 2021	Exploratory Data Analysis
10	01 Mar 2021	Statistical Inference
11	08 Mar 2021	Statistical Modeling I
12	15 Mar 2021	Statistical Modeling II
13	22 Mar 2021	Data Visualization
14	29 Mar 2021	Data Communication

Motivation



Reproducible data analysis

- Data analysis can be repeated by others (transparent and reproducible)
- Analogous to lab protocol
- Journal requirement to submit data and code

Avoid using Excel!

- Difficult to separate the data from the process
- Difficult to follow the logic behind the analysis
- Formulas are hidden in cells, can be accidentally deleted or overwritten
- Prone to formatting errors

Avoid using Excel!



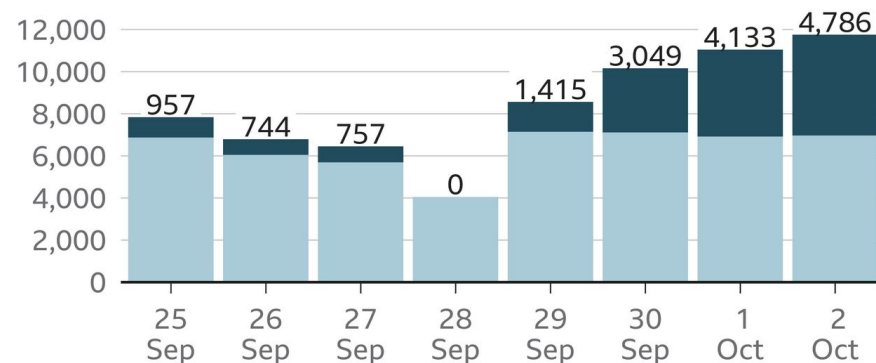
Covid: Test error 'should never have happened' - Hancock

5 October

Thousands of missing coronavirus cases added after reporting problem

Number of new coronavirus cases by date reported

■ Missing cases added ■ Previously announced cases



Source: Gov.uk dashboard, Public Health England

BBC

Avoid using Excel!

Ziemann et al. *Genome Biology* (2016) 17:177
DOI 10.1186/s13059-016-1044-7

Genome Biology

COMMENT

Open Access

Gene name errors are widespread in the scientific literature

Mark Ziemann¹, Yotam Eren^{1,2} and Assam El-Osta^{1,3*}



Abstract

The spreadsheet software Microsoft Excel, when used with default settings, is known to convert gene names to dates and floating-point numbers. A programmatic scan of leading genomics journals reveals that approximately one-fifth of papers with supplementary Excel gene lists contain erroneous gene name conversions.

Keywords: Microsoft Excel, Gene symbol, Supplementary data

Abbreviations: GEO, Gene Expression Omnibus; JIF, journal impact factor

The problem of Excel software (Microsoft Corp., Redmond, WA, USA) inadvertently converting gene symbols to dates and floating-point numbers was originally described in 2004 [1]. For example, gene symbols such as *SEPT2* (Septin 2) and *MARCH1* [Membrane-Associated Ring Finger (C3HC4) 1, E3 Ubiquitin Protein Ligase] are converted by default to '2-Sep' and '1-Mar', respectively. Furthermore, RIKEN identifiers were described to be automatically converted to floating point numbers (i.e. from accession '2310009E13' to '2.31E+13'). Since

frequently reused. Our aim here is to raise awareness of the problem.

We downloaded and screened supplementary files from 18 journals published between 2005 and 2015 using a suite of shell scripts. Excel files (.xls and .xlsx suffixes) were converted to tabular separated files (tsv) with *ssconvert* (v1.12.9). Each sheet within the Excel file was converted to a separate tsv file. Each column of data in the tsv file was screened for the presence of gene symbols. If the first 20 rows of a column contained five or more gene symbols, then it was suspected to be a list of gene symbols, and then a regular expression (regex) search of the entire column was applied to identify gene symbol errors. Official gene symbols from Ensembl version 82, accessed November 2015, were obtained for *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Danio rerio*, *Escherichia coli*, *Gallus gallus*, *Homo sapiens*, *Mus musculus*, *Oryza sativa* and *Saccharomyces cerevisiae* [2]. The regex search used was similar to that described previously by Zeeberg and colleagues [1], with the added screen for dates in other formats (e.g. DD/MM/YY and MM-DD-YY). To expedite analysis of supplementary files from multi-disciplinary journals, we limited the articles screened to those that have the keyword 'genome' in the title or abstract (*Science*, *Nature* and *PLoS One*). Excel files (.xls and .xlsx) deposited

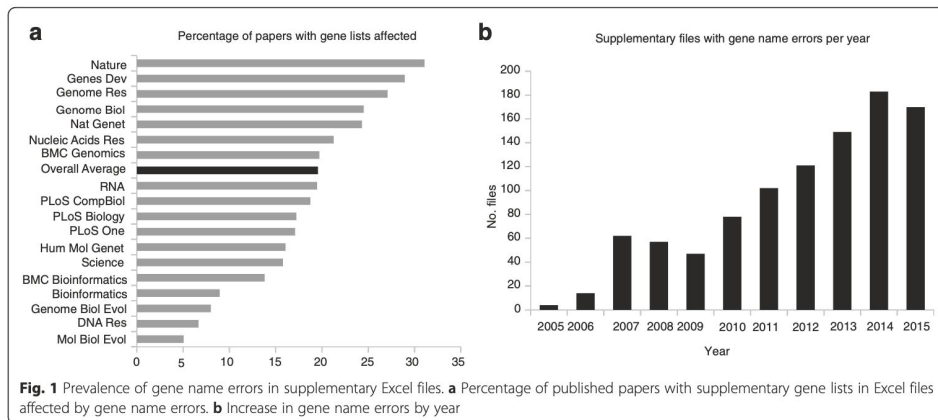
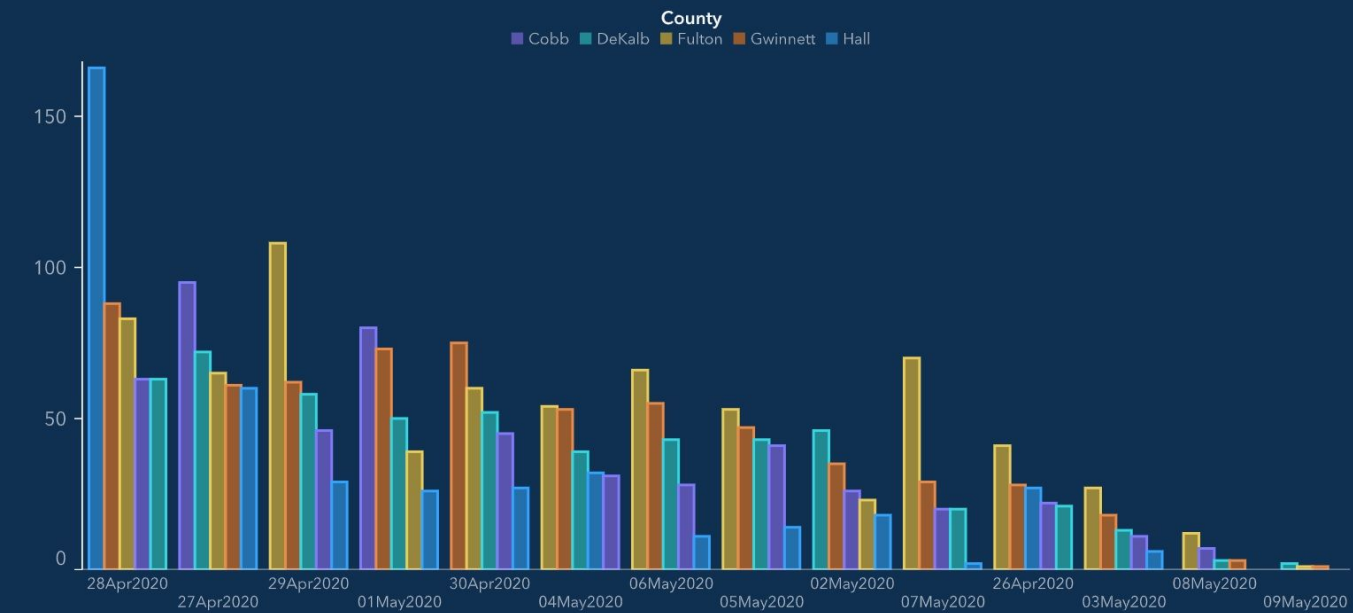


Fig. 1 Prevalence of gene name errors in supplementary Excel files. **a** Percentage of published papers with supplementary gene lists in Excel files affected by gene name errors. **b** Increase in gene name errors by year

Data analysis gone wrong !

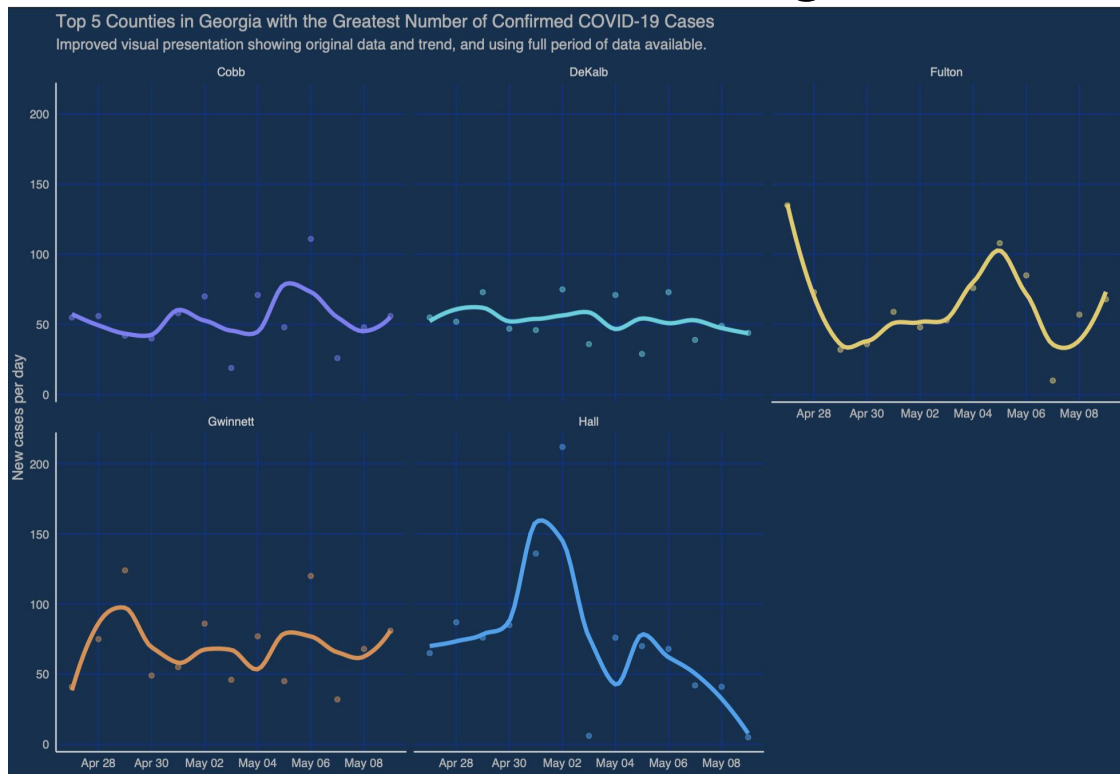
Top 5 Counties with the Greatest Number of Confirmed COVID-19 Cases

The chart below represents the most impacted counties over the past 15 days and the number of cases over time. The table below also represents the number of deaths and hospitalizations in each of those impacted counties.



Department of Public Health,
Georgia, USA

Better visualization using “small multiples”



<https://www.r-bloggers.com/2020/05/ordering-bars-within-their-clumps-in-a-bar-chart-by-ellis2013nz/>

Why R?



- Free software and open-source
- Environment for statistical computing and graphics
- Simple programming language based on S (1976)
- Comprehensive set packages from accessing data, cleaning, analysis, and reporting
- Community of developers
- Easy to install packages

A screenshot of the R Console window on a macOS system. The window has a title bar with standard macOS window controls (red, yellow, green buttons) and a title 'R Console'. Below the title bar is a toolbar with icons for a stop button, R logo, a bar chart, a lock, a flag, a color wheel, a document with the R logo, a plain document, a printer, and a mobile device. A search bar with the text 'Help Search' is also present. The main content area displays the R startup message for version 4.0.3 (2020-10-10) on a 64-bit Darwin platform. It includes the copyright notice for The R Foundation, a warning about no warranty, and instructions on how to use various help functions like 'license()', 'contributors()', 'citation()', 'demo()', and 'help.start()'. It also shows the GUI version and workspace/history restoration status. The prompt '>' is visible at the bottom.

```
R version 4.0.3 (2020-10-10) -- "Bunny-Wunnies Freak Out"
Copyright (C) 2020 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin17.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

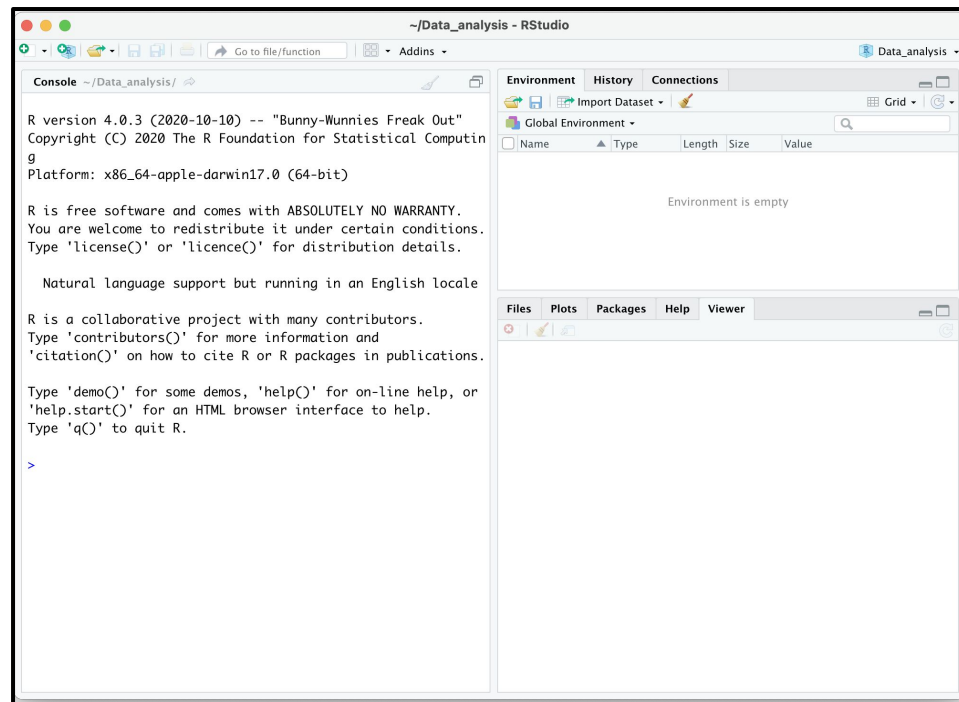
[R.app GUI 1.73 (7892) x86_64-apple-darwin17.0]

[Workspace restored from /Users/clauidius/.RData]
[History restored from /Users/clauidius/.Rapp.history]

> |
```

R under GUI using RStudio

- Easy to use interface
- Console is where you can type commands and see output
- Workspace tab shows all active objects
- History tab shows a list of commands recently used
- Files tab shows all files and folders in your workspace
- Plots tab show all graphs
- Packages tab will list of series of packages
- Help tab to see additional information



How R works?

- R creates objects in memory and saves them in a file called .RData
- Commands are recorded in an .Rhistory file (recall by pressing arrow up or down)
- Recalled commands may be edited
- Commands may be abandoned by pressing <Esc>
- To end your session, type q() or just kill the window
- Use of **working directory**: each project is associated with a working folder containing all data, scripts, output files, figures, etc.

Hands-on exercises using R

- Focus on real-life data, e.g. COVID-19 dataset at Johns Hopkins University
- Clean messy data
- Explore data
- Visualize data
- Do basic statistical analyses
- Generate models
- Generate publication ready figures
- Communicate results

Using R for documentation

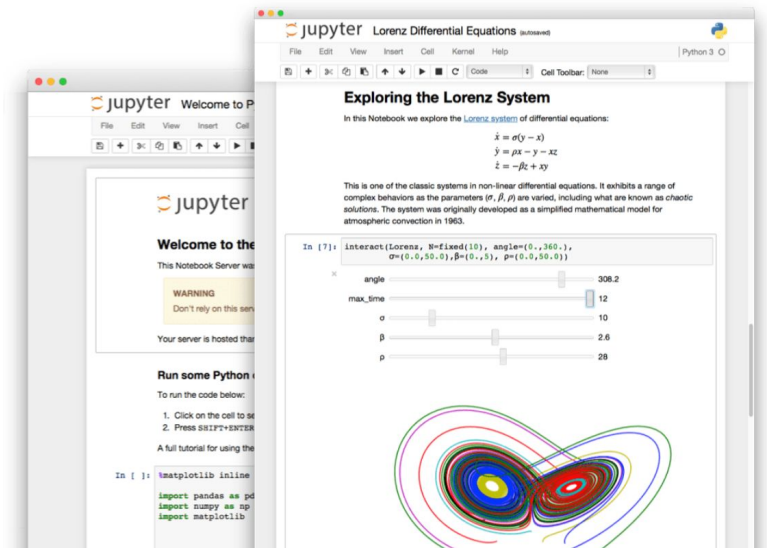
- virtual “lab notebook”
- R Script, R Markdown, R Notebook, Jupyter
- For capturing what you did, codes, results, figures, conclusions
- For collaborating with other team members, your boss, future you!
- For submitting to journals as part of publication (reproducibility)

Using Jupyter

- For hands-on exercises in this course



[Install](#) [About Us](#) [Community](#) [Documentation](#) [NBViewer](#) [JupyterHub](#) [Widgets](#) [Blog](#)



The Jupyter Notebook

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

Try it in your browser

Install the Notebook

What is not covered in this course?

- Crafting a research question
- Gathering data
- Teaching computer to learn from data

Google Classroom

☰ Data Analysis

Stream

Classwork

People

Grades



Data Analysis

Class code ssmnn2n

Meet link <https://meet.google.com/lookup/b3bmxetfn>

Select theme
Upload photo

Upcoming

No work due soon

[View all](#)



Share something with your class...



CLYDE PANCITO DAPAT

Nov 8 (Edited Nov 8)



Hi! Welcome to the Data Analysis course.
Classes will start on December 7, 2020 at 10 AM.

Google Classroom

≡ Data Analysis

Stream

Classwork

People

Grades



+ Create



Meet



Google Calendar



Class Drive folder

All topics

📖 Course Resources

👋 1. Introduction to...

💻 2. From Variables...

🔧 3. Control Flow (...)



Course Resources



Class Syllabus and Schedule

Posted Nov 8



1. Introduction to Data Analysis (07 Dec 202... :

Google Classroom

☰ Data Analysis

Stream

Classwork

People

Grades



Teachers



CLYDE PANCITODAPAT

Students

7 students



Actions ▾

A-Z



MATEJOVIC ADAM



FANG CHEN



MALAGA GRANDA ...



Course Project

- Describe your project question
- Plan in answering your project question
- Present code, figures, results, and conclusion
- Class presentation at end of course

Things to do

- Install R, RStudio, Jupyter and R kernel (installation guide available in Classwork tab of Google Classroom)
- Take the course survey

Take away message

- Use R for data analysis
- Data analysis should be reproducible