

# Data Analysis Workflow .. and some other stuffs

Kamarul Imran Musa

Associate Professor (Epidemiology and  
Statistics)

Public Health Physician

Fellow of the American College of  
Epidemiology





# Biography

- MD (USM), Master of Community Medicine and PhD in Epidemiology and Statistics from Lancaster University
- Research interests:
  - Epidemiological and Statistical Modelling of Diseases
  - Predictive Analysis using Machine Learning (including Deep Learning)
- Principal investigator for
  - Mal-UK Stroke Study (Newton)
  - Modelling for breast cancer (FRGS)
  - CVD diseases and risk factors modelling
- Fellow of the American College of Epidemiology
- <https://myanalytics.com.my/>
- <https://medic.usm.my/jpm/index.php/en/academic-information/587-prof-madya-dr-kamarul-imran-musa>

# R



- R is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories
- R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, ...) and graphical techniques, and is highly extensible.
- One of R's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed.
- R is available as Free Software under the terms of the Free Software Foundation's GNU General Public License in source code form.

A large orange circle is positioned on the left side of the slide, partially cut off by the edge.

Link to slides

- <https://bit.ly/Rppsp21>



# RStudio

RStudio is an integrated development environment (IDE) for R. It includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, debugging and workspace management.

RStudio is available in open source and commercial editions and runs on the desktop (Windows, Mac, and Linux) or in a browser connected to RStudio Server or RStudio Workbench (Debian/Ubuntu, Red Hat/CentOS, and SUSE Linux).



# R at USM



## Account Information

Account Name University Sains Malaysia (USM)  
Contact Name Kamarul Imran Musa  
Phone  
Email drkamarul@usm.my

Product Code & Description	Quantity	Subscription Term (Months)	Total Subscription Fee
RSC-Base	1.00	12.00	\$14,995.00
RStudio Connect Base (Single Server) with 20 Named Users			
RSW-Standard-Named-User	22.00	12.00	\$20,900.00
RStudio Workbench Standard Named User (Single Server). See <a href="https://rstudio.com/about/software-license-descriptions/">https://rstudio.com/about/software-license-descriptions/</a> for complete information.			
Order Total:			\$35,895.00



name and password, please request from drkamarul[at]usm.my

# R in Medicine

## R/Medicine Virtual Conference

August 24th - 27th 2021

[View Schedule and Recordings](#)

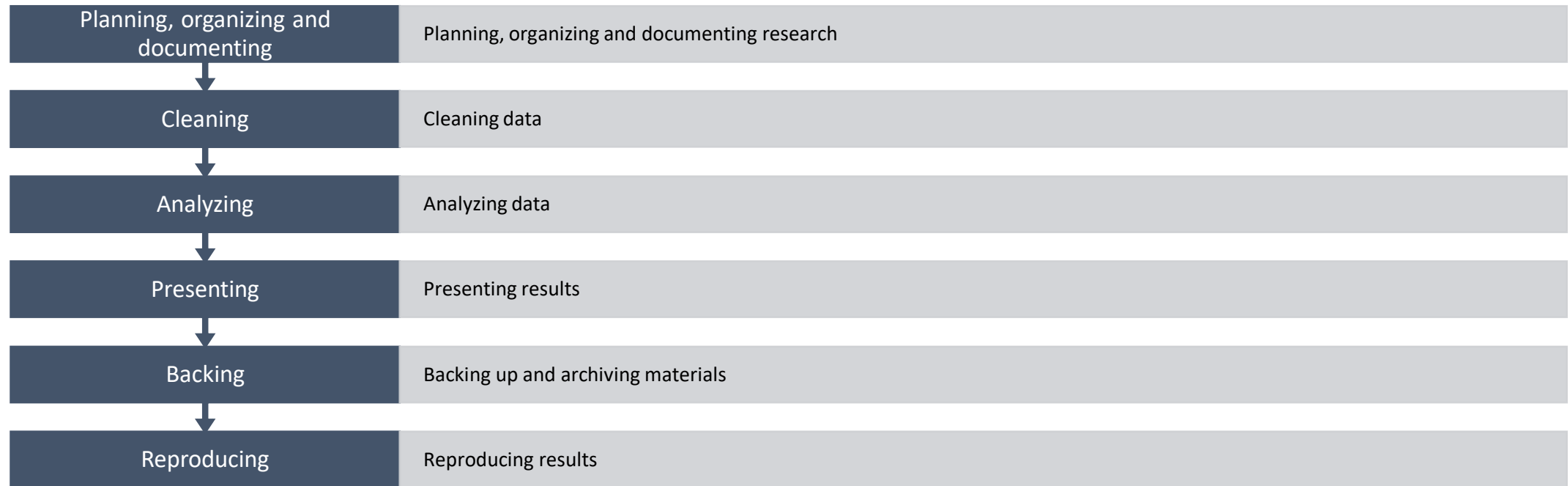
[View Posters](#)



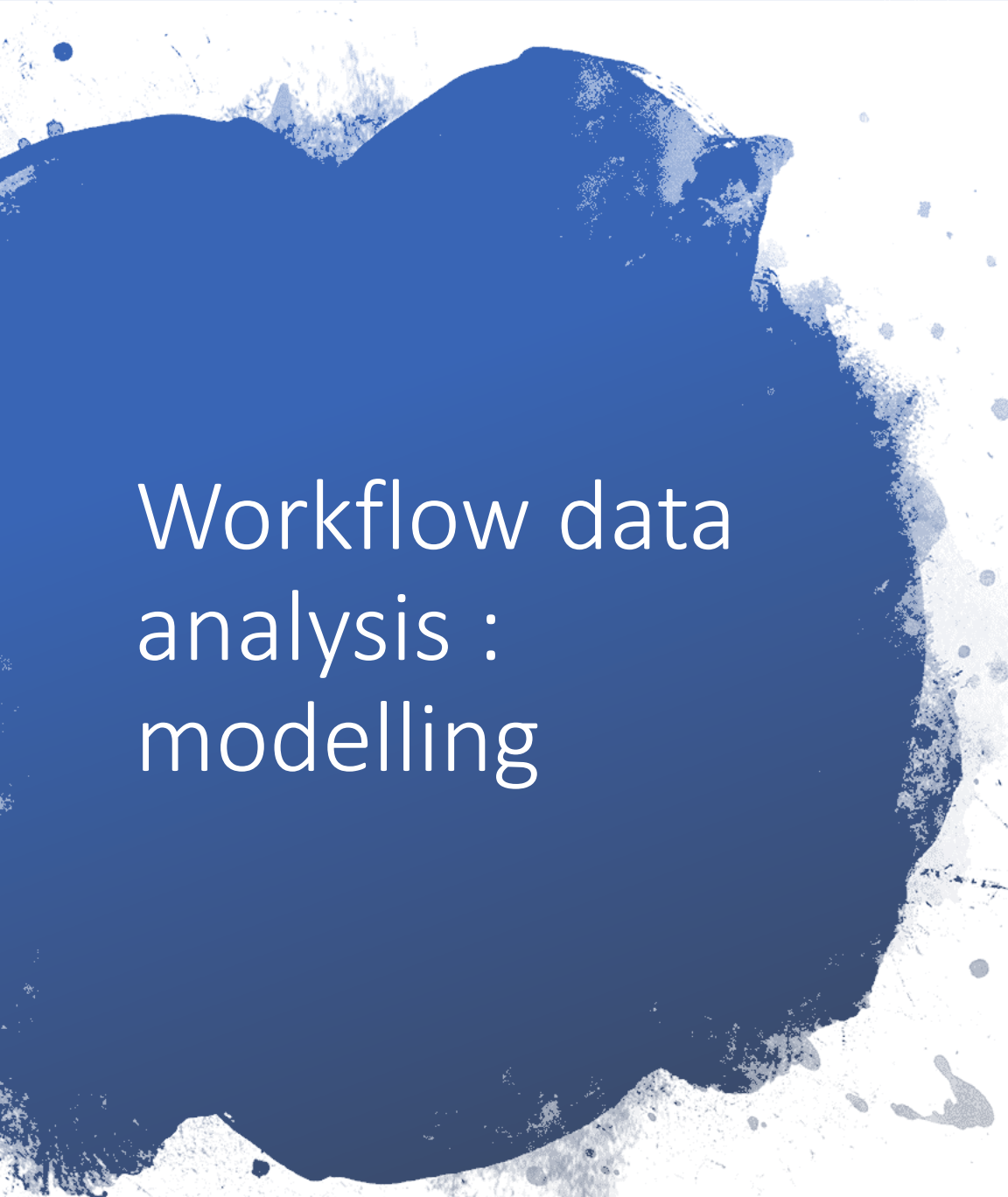
**BROUGHT TO YOU BY**



# Workflow for data analysis : stages



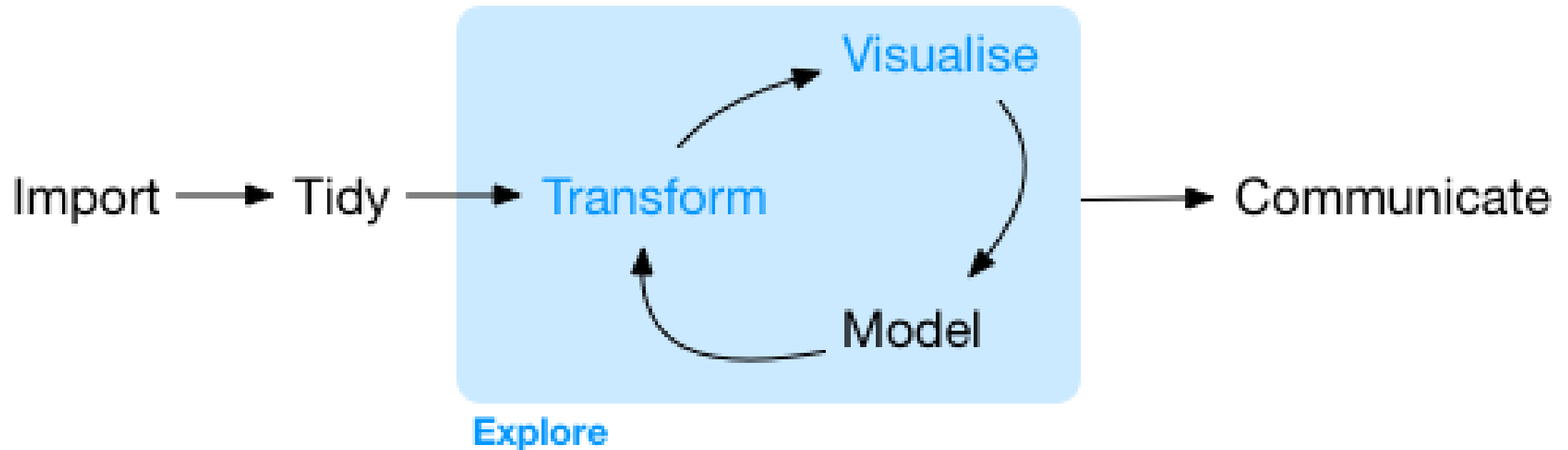




# Workflow data analysis : modelling

- Read data
- Clean data
- Describe data
- Model data
- Check model
- Interpret model
- Communicate model

# Workflow data analysis : tools



Program

<http://r4ds.had.co.nz/introduction.html>

# Tidy

- Tidying your data means storing it in a consistent form that matches the semantics of the dataset with the way it is stored.
- In brief, when your data is tidy, each column is a variable, and each row is an observation.
- Tidy data is important because the consistent structure lets you focus your struggle on questions about the data, not fighting to get the data into the right form for different functions.

# Transform

- Transformation includes narrowing in on observations of interest (like all people in one city, or all data from the last year), creating new variables that are functions of existing variables (like computing velocity from speed and time), and calculating a set of summary statistics (like counts or means). Together, tidying and transforming are called wrangling,

# Visualization

- A good visualisation will show you things that you did not expect, or raise new questions about the data.
- A good visualisation might also hint that you're asking the wrong question, or you need to collect different data.



# Tidyverse

## Tidyverse



### R packages for data science

The tidyverse is an opinionated **collection of R packages** designed for data science. All packages share an underlying design philosophy, grammar, and data structures.

Install the complete tidyverse with:

```
install.packages("tidyverse")
```

# Models

- Models are a fundamentally mathematical or computational tool, so they generally scale well.
- Even when they don't, it's usually cheaper to buy more computers than it is to buy more brains!
- But every model makes assumptions, and by its very nature a model cannot question its own assumptions.

# Communication

- It doesn't matter how well your models and visualisation have led you to understand the data unless you can also communicate your results to others.
- Where
  - Outside organization
  - Within organization

# Programming

- Programming is a cross-cutting tool that you use in every part of the project.
- You don't need to be an expert programmer to be a data scientist, but learning more about programming pays off

# Communicate

testrstudioconnect

Kamarul Imran Musa  
9/5/2018

## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

##	speed	dist
## Min.	: 4.0	Min. : 2.00
## 1st Qu.:	12.0	1st Qu.: 26.00
## Median :	15.0	Median : 36.00
## Mean :	15.4	Mean : 42.98
## 3rd Qu.:	19.0	3rd Qu.: 56.00
## Max. :	25.0	Max. : 120.00

## RStudio Connect

## Including Plots

You can also embed plots, for example:

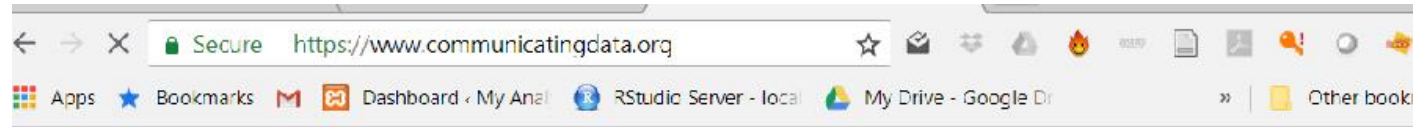
RStudio Connect is a new publishing platform for the work your teams create in R. Share Shiny applications, R Markdown reports, Plumber APIs, dashboards, plots, and more in one convenient place. Use push-button publishing from the RStudio IDE, scheduled execution of reports, and flexible security policies to bring the power of data science to your entire enterprise.

TRY THE FREE 45 DAY EVALUATION

SCHEDULE A MEETING WITH SALES



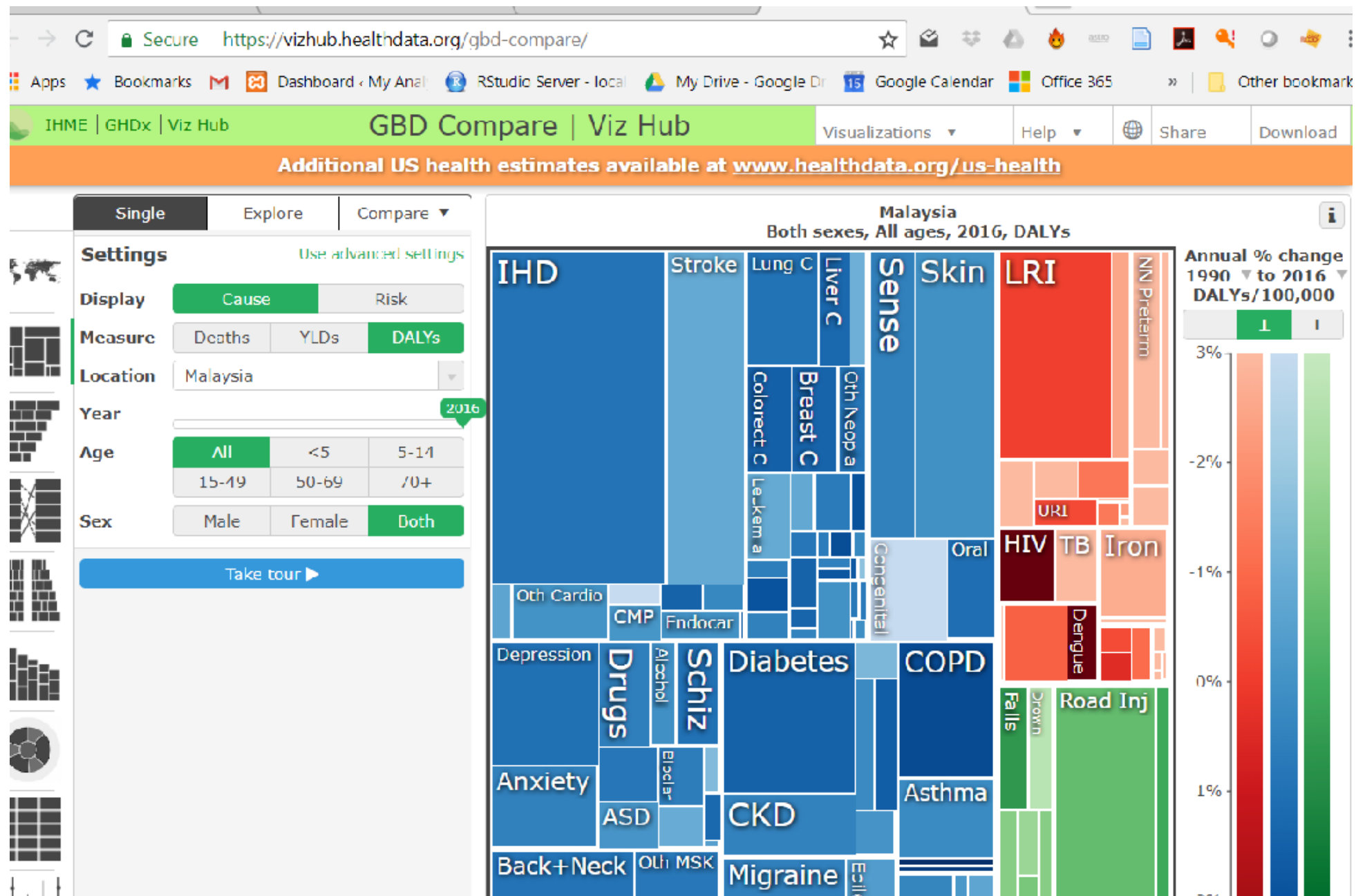
# Communicate



Communicating  
Data for Impact

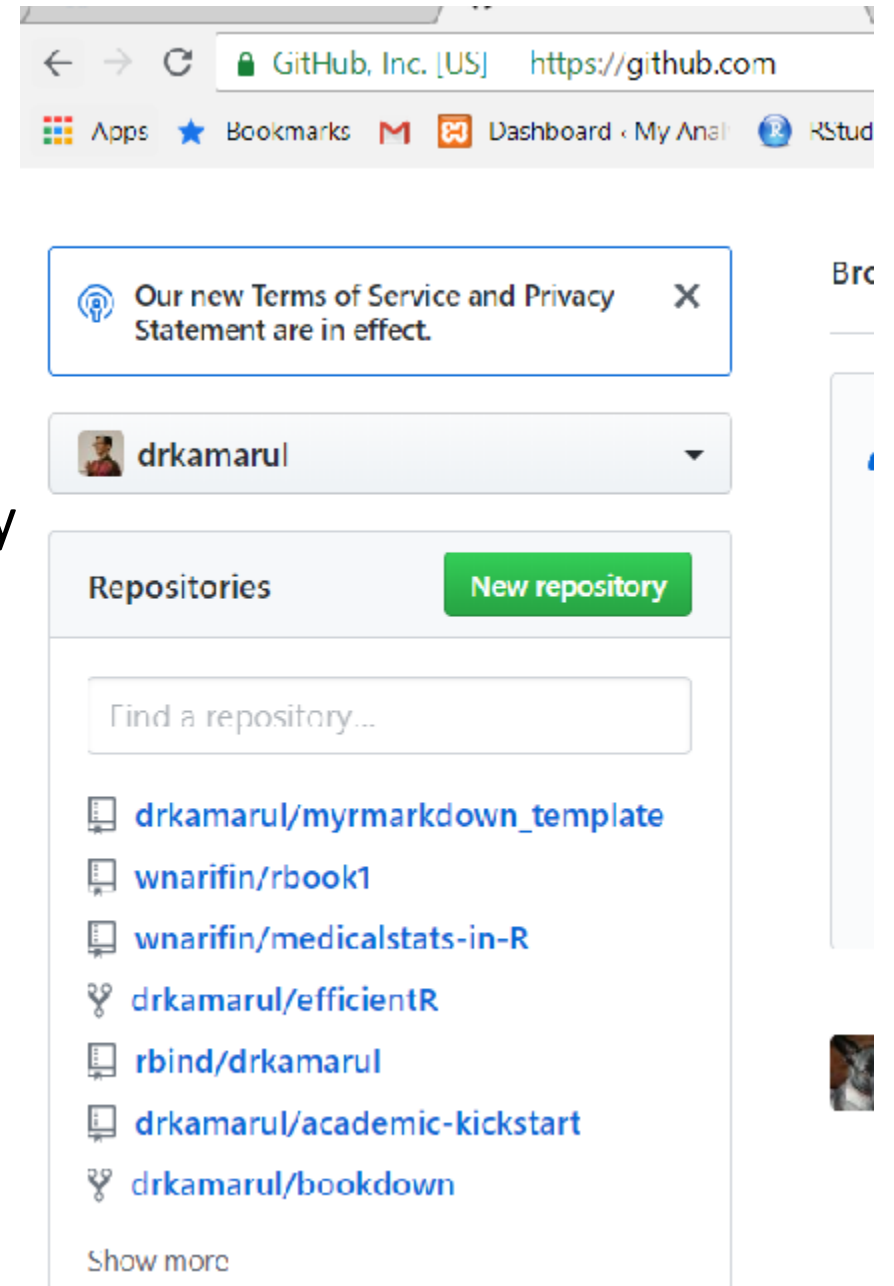
[Home](#) [Case Study](#) [White Paper](#) [About](#)





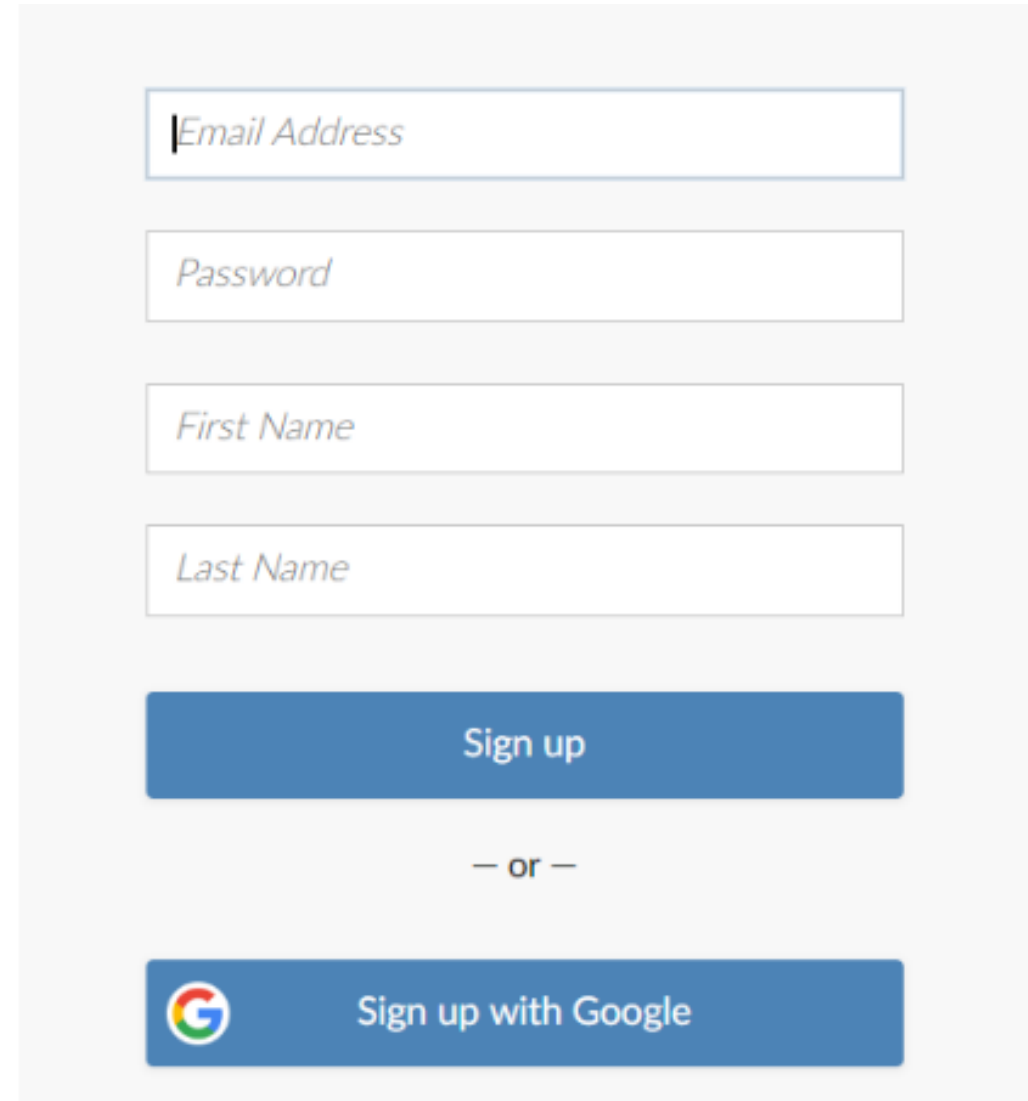
- Version control
  - Why use git?
    - Version control is the only reasonable way to keep track of changes in code, manuscripts, presentations, and data analysis projects
  - Why use [github](https://github.com)?
    - *Github is like facebook for programmers.* Everyone's on there. You can look at what they're working on and easily peruse their code and make suggestions or changes.

[http://kbroman.org/github\\_tutorial/pages/why.html](http://kbroman.org/github_tutorial/pages/why.html)



# Sign-up for Rstudio Cloud

- Sign-up with RStudio Cloud here  
[https://client.login.rstudio.cloud/oauth/register?redirect=https%3A%2F%2Fclient.login.rstudio.cloud%2Foauth%2Flogin%3Fshow\\_auth%3D0%26show\\_login%3D1%26show\\_setup%3D1](https://client.login.rstudio.cloud/oauth/register?redirect=https%3A%2F%2Fclient.login.rstudio.cloud%2Foauth%2Flogin%3Fshow_auth%3D0%26show_login%3D1%26show_setup%3D1)



The image shows a sign-up form for RStudio Cloud. It consists of four text input fields stacked vertically: 'Email Address', 'Password', 'First Name', and 'Last Name'. Below these fields is a blue button labeled 'Sign up'. Underneath the button is the text '— or —'. At the bottom is another blue button featuring the Google logo and the text 'Sign up with Google'.

Email Address


Password

First Name

Last Name

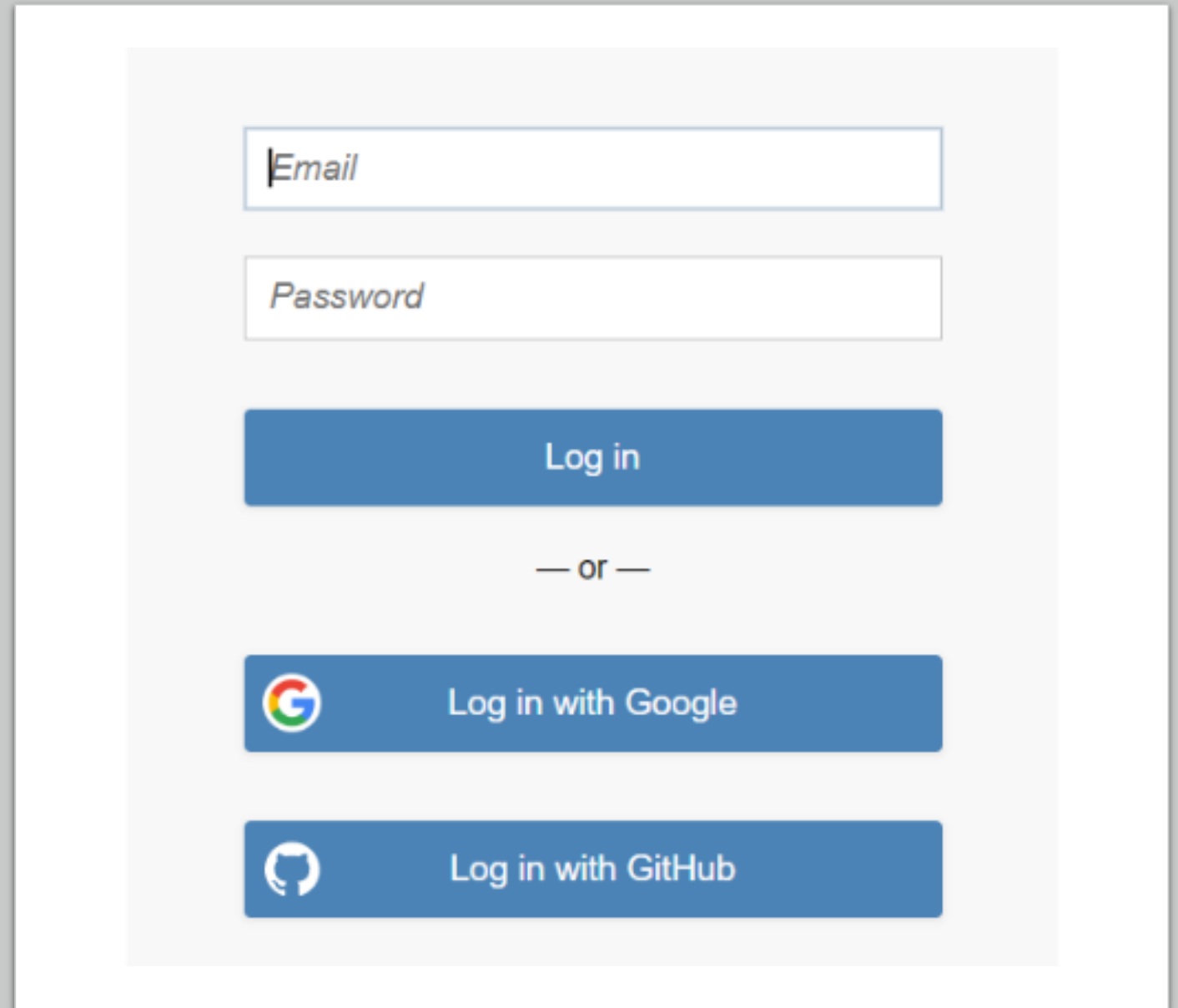
Sign up

— or —

 Sign up with Google

# Login to Rstudio Cloud

- Login to RStudio Cloud




The image shows the login interface for RStudio Cloud. It features a light gray background with a white border. At the top, there is a text input field labeled "Email" with a cursor inside. Below it is a text input field labeled "Password". Underneath the password field is a blue button with the text "Log in". Below the "Log in" button is the text "— or —". Below this text are two more blue buttons. The first button has the Google logo on the left and the text "Log in with Google" on the right. The second button has the GitHub logo on the left and the text "Log in with GitHub" on the right.


Email

Password

Log in

— or —

 Log in with Google

 Log in with GitHub



# Link to our shared project

- [https://rstudio.cloud/spaces/170611/join?access\\_code=%2FrGl%2BDmMyfqTx268w8ZzA5n1JrFyJv%2FYNqNilm2N](https://rstudio.cloud/spaces/170611/join?access_code=%2FrGl%2BDmMyfqTx268w8ZzA5n1JrFyJv%2FYNqNilm2N)

## Join Space?

Joining a space gives you access to it and to its contents.

Once you join, admins will be able to see your email address.

Would you like to join this space?

Join Space

## Welcome to Workshop on R - PPSP

If you did not intend to join this space, or you later decide you don't want to be a member, just go to the [Members](#) area and click "Leave Space".

- Klik PROJECTS
- Click START

Workshop on R - PPSP  
Kamarul Imran (KIM) M



Projects Members About

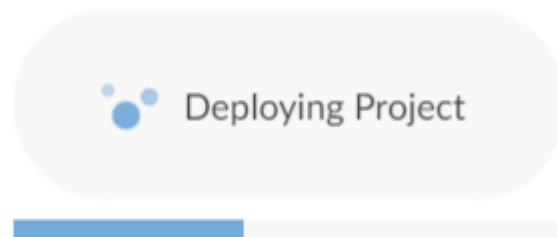
All Projects

List  Sort

**START** Introduction (Session 1)

R workshop for PPSP

 Kamarul Imran (KIM) M  RStudio Project Created Sep 14, 2021 11:10 PM



Workshop On R - PPSP / Introduction (Session 1)

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

Console Terminal x Jobs x

R 4.1.0 . /cloud/project/

```
R version 4.1.0 (2021-05-18) -- "Camp Pontanezen"
Copyright (C) 2021 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (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.

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.

> |
```

Environment History Conne

R Global Environment

Files Plots Packages Help

New Folder Upload

Cloud > project

Name

- ..
- .Rhistory
- Data\_Visualization.Rmd
- Intro\_to\_R.Rmd
- project.Rproj

# Skills

At the end of each session, rate yourself from 0-10 (0 is like totally clueless and 10 means you can teach others)

R = Read data

E = Explore data

A = Analyze data

P = Present findings

# Ingredients

Package – the right library that contains a set of functions to do desired task

Function – contains codes that you type so R will perform the desired task

Data – that you want to analyze

Make plots

Ingredient	
library	tidyverse gapminder
functions	ggplot
dataset	gapminder peptic ulcer stroke



# References

- <http://r4ds.had.co.nz/index.html>

