Rotman

INTRO TO R

R Workshop – Part 1 Overview & Basics / 1

March 13, 2024 Prepared by Jay Cao / TDMDAL

Website: https://tdmdal.github.io/r-workshop-202324-winter/

Rotman School of Management
UNIVERSITY OF TORONTO

Plan for the 4-Session Workshop

- Part 1: Overview & Basics (Session 1, 2)

- Part 2: Data Manipulation (Session 2, 3)

- Part 3: Data Visualization (Session 3)

- Part 4 - 1: Modeling Workflow (Session 4)
- Part 4 - 2: Time Series & Some Finance Applications (Session 4)

Plan for Part 1

- Intro
  - What is R and what can R do?
  - Setup R
  - Motivation examples

- R programing and Data Science
  - Basics of R programming
  - Data science with R

- Learning Resources and Road Map

What’s R?

- R = a language + an eco-system
  - A free and open-source programming language
  - An eco-system of many high-quality user-contributed libraries/packages

- In the past R is mostly known for its statistical analysis toolkits

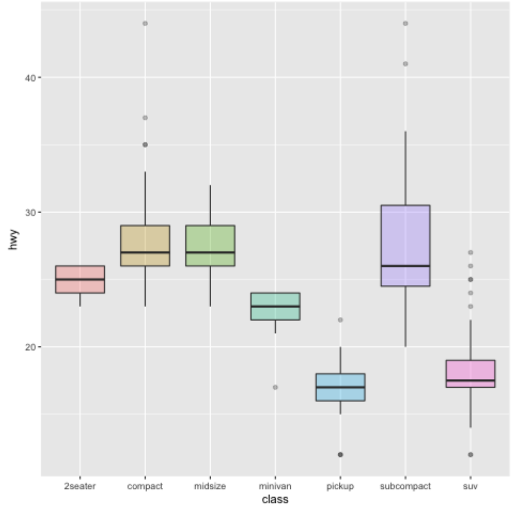
- Nowadays R is capable of (and very good at) many other tasks
  - Tools that facilitate the whole data analysis workflow
  - Tools for web technology
  - Many more…

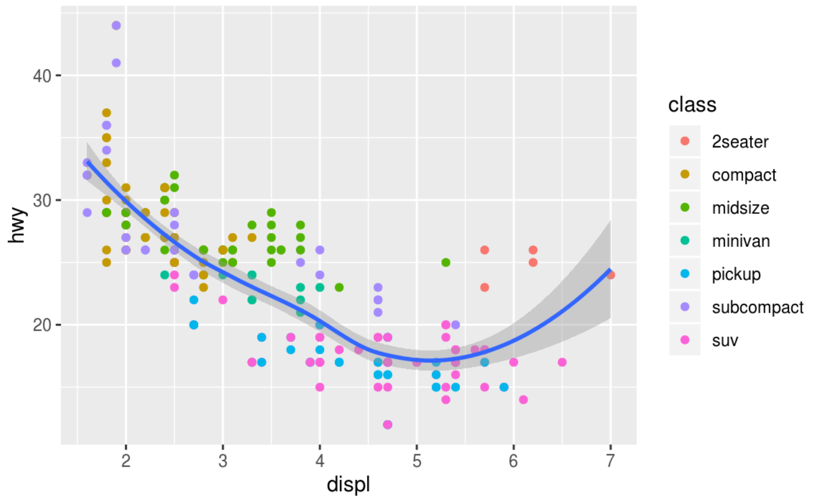
What can R do – Statistics & related

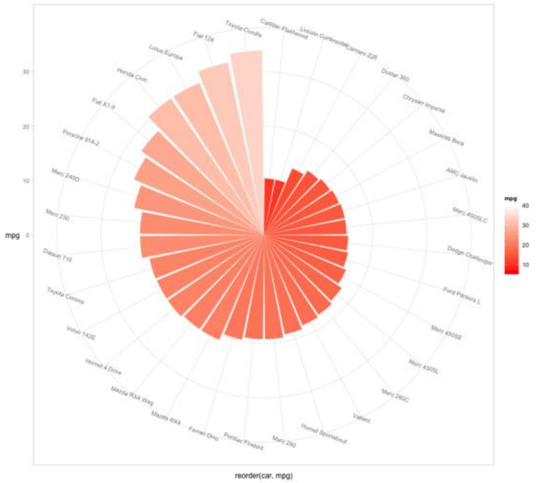
- Statistics & Econometrics
  - Regressions
  - Time series analysis
  - Bayesian inference
  - Survival analysis
  - …
- Numerical Mathematics
  - Optimization
  - Solver
  - Differential equations
  - …
- Finance
  - Portfolio management
  - Risk management
  - Option pricing
  - …
- …

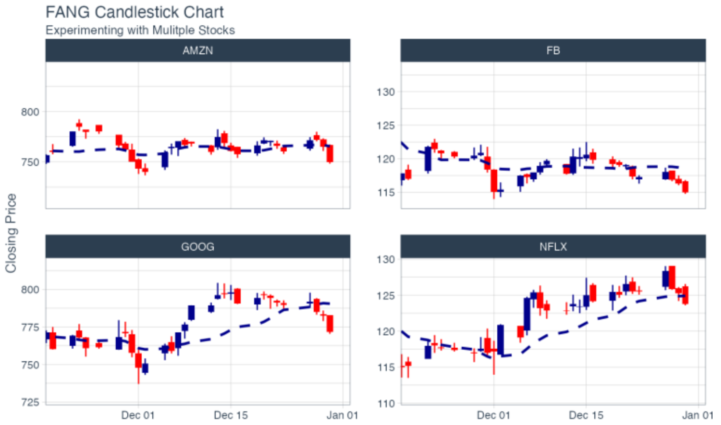
See more R Packages on R Task View and more Empirical Finance Packages on R Task View - Finance

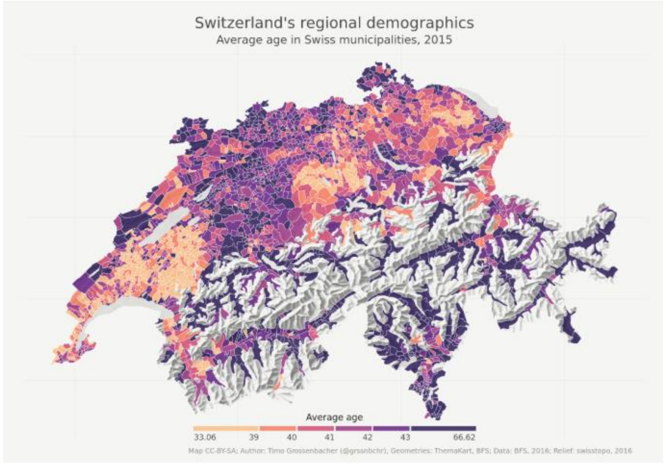
What can R do – Graphics (static ones)

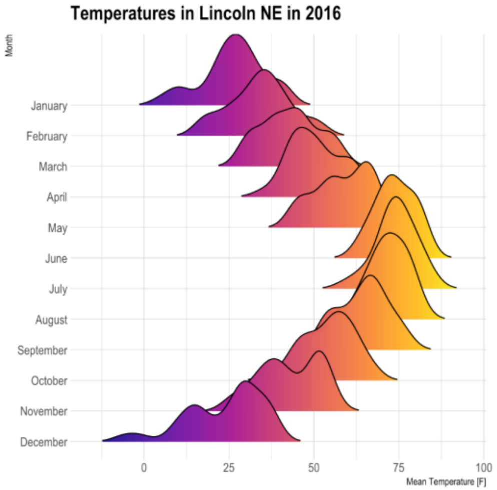









Switzerland's regional demographics
Average age in Swiss municipalities, 2015
Average age
33.06
39
40
41
42
43
66.62

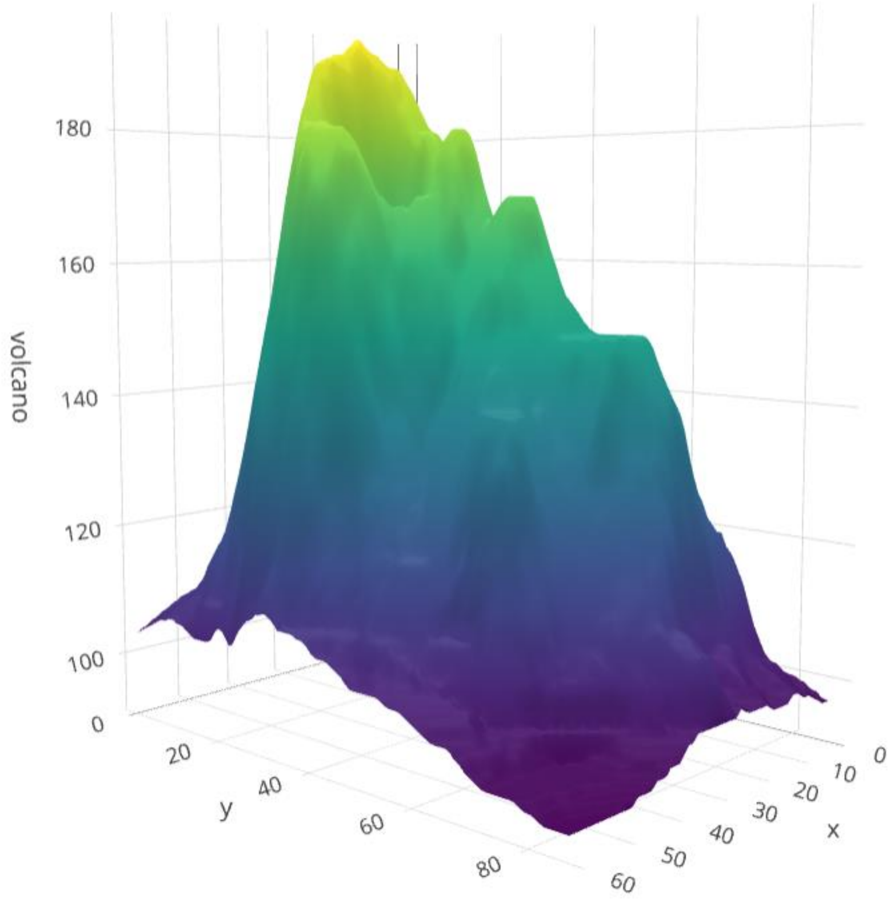


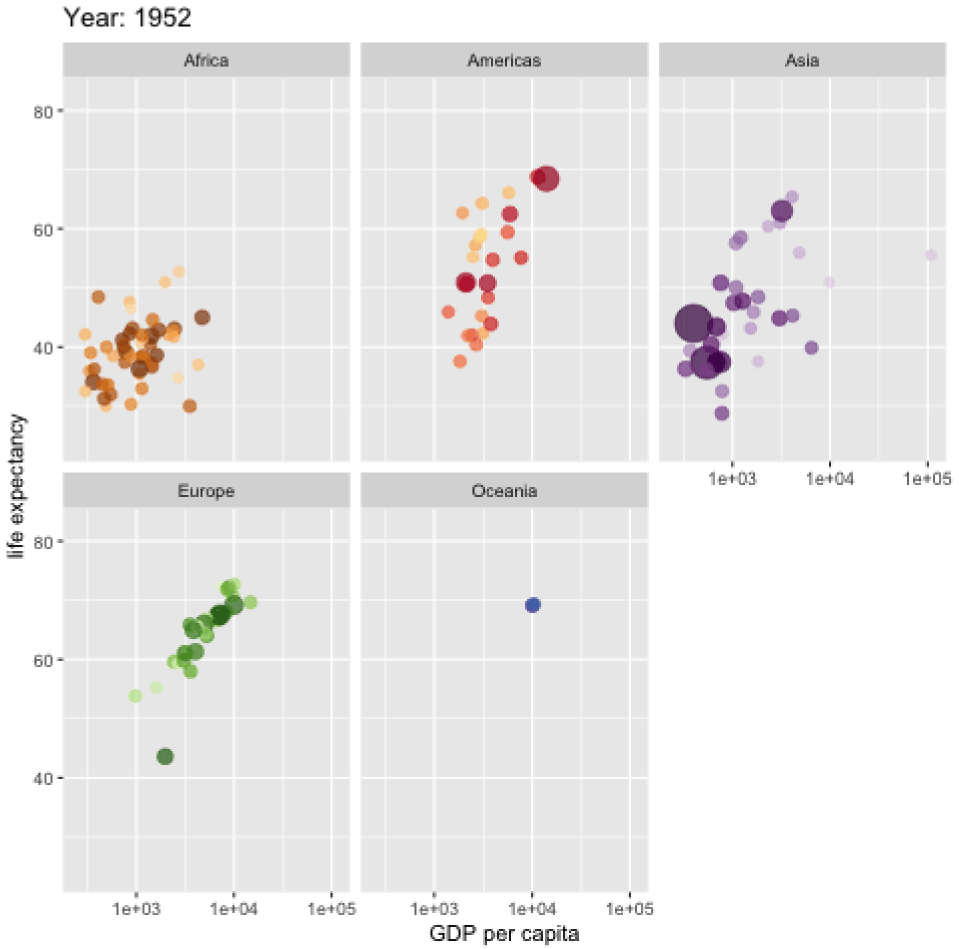
...

Ref. 1) https://www.r-graph-gallery.com/

2) https://timogrossenbacher.ch/bivariate-maps-with-ggplot2-and-sf/

What can R do – Graphics (dynamic ones)

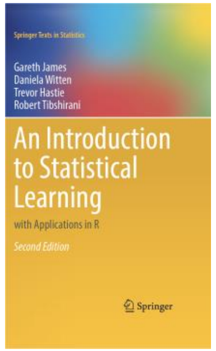




https://plot.ly/r/3d-surface-plots/;

https://gganimate.com/;

What can R do – ML & NLP



Springer Texts in Statistics

Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani

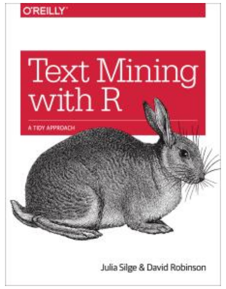
An Introduction to Statistical Learning

with Applications in R

Second Edition

Springer

- Machine learning
  - Statistical learning (clustering, decision tree, etc.)
    - An Introduction to Statistical Learning (with Applications in R)
  - Deep learning (neural networks)
    - Tensorflow for R (via reticulate, an R to Python interface)
    - Torch for R (natively from R; similar as PyTorch in Python)
- Natural language processing
  - Packages (e.g., tidytext, topicmodels)
  - Books (e.g., Text Mining with R, Supervised ML for Text Analysis in R)
  - Leveraging the Python Transformers library (e.g., Transformers from R)



O'REILLY

Text Mining with R

A TIDY APPROACH

Julia Silge & David Robinson

DATA SCIENCE SERIES

SUPERVISED MACHINE LEARNING FOR TEXT ANALYSIS IN R

EMIL HVITFELDT
JULIA SILGE

CRC Press

1. See more R Machine Learning Packages on R Task View - ML & Statistical Learning
2. See more R Natural Language Processing Packages on R Task View - NLP

What can R do – Web & Reporting

- Web technology
  - Web scraping (e.g., rvest)
  - API wrapper (e.g., Twitter: rtweet; bigquery: bigrquery; Quandl: Quandl)
  - Shiny web app (https://shiny.rstudio.com/)

- Reporting
  - R Markdown (write reports, slides, blogs, books, etc. See a gallery here.)
  - Quarto (new authoring tool; multi-language and multi-engine;)

- … (see **R Task View** for more)

R vs Excel and BI Tools vs Python

- Excel & Business Intelligence (BI) Tools (e.g., Tableau, Power BI, etc.)
  - 2-D tables as basic data structure
  - Good UI (User Interface) and minimum programming
  - Limited modeling tools
  - Not easy to reproduce an analysis (because it’s hard to store UI clicks)
  - Not flexible enough for complicated analytics problems, i.e., problems with
    - Many data cleaning steps/pipelines
    - Many different models to try

- Python
  - Python is more general purpose; R is more specialized in statistical analysis
  - R is much easier to learn (in my opinion)

Why learn R (What can R do **for You**)?

- Beyond Excel Data Analysis
  - I wish Excel could…
- Automate boring repeating tasks
  - e.g., daily data collection from different sources, weekly dashboard update
- Prototype ideas
  - e.g., a novel trading strategy, a new credit risk model
- Really, find anything that interests you and use R…

Plan for Part 1

- Intro
  - What is R and what can R do?
  - Setup R
  - Motivation examples

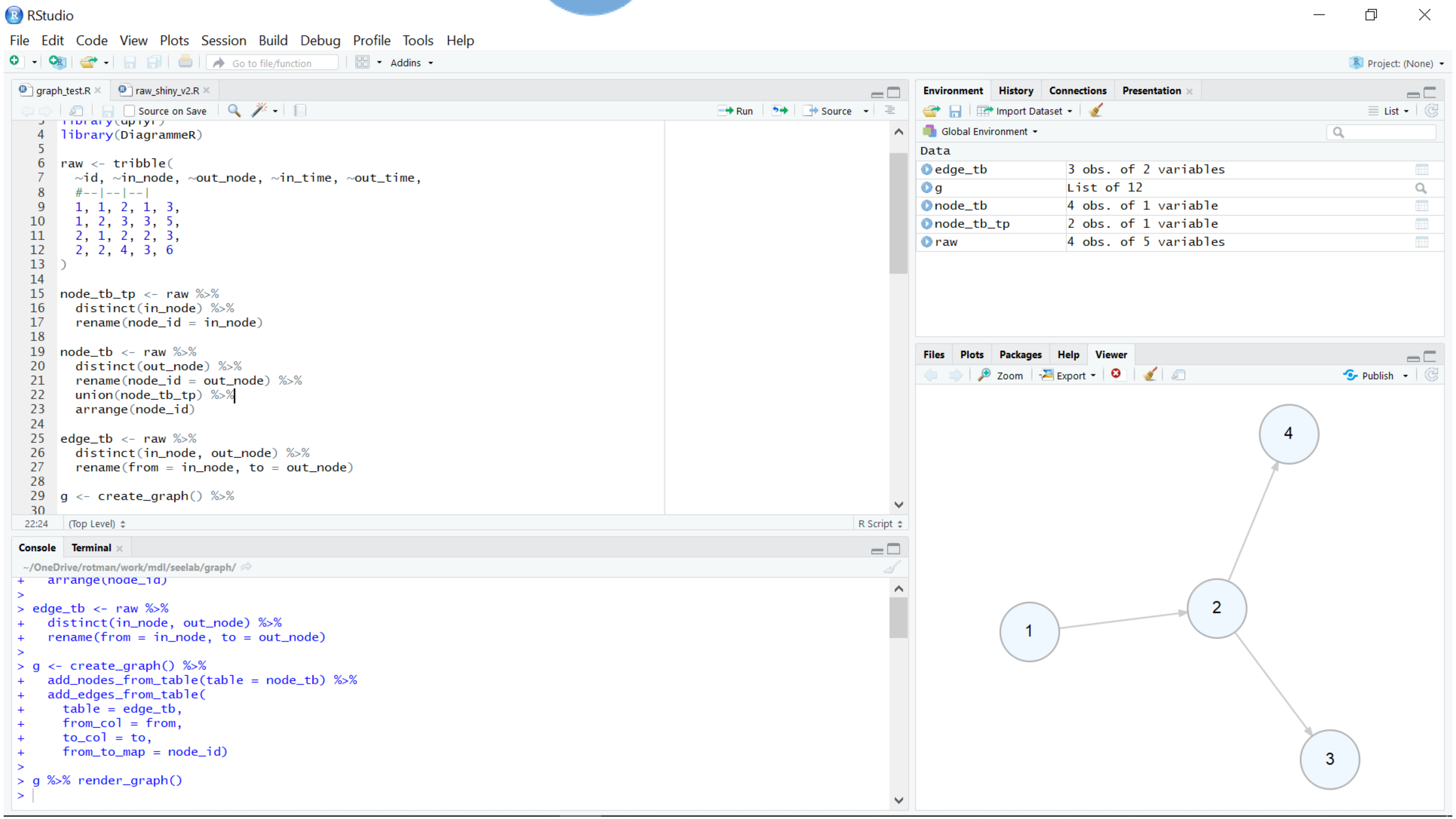
- Overview of R programing and Data Science
  - Basics of R programming
  - Data science with R

- Learning Resources and Road Map

Setup R (Install R & its Coding Environment)

	R & RStudio	**R & Notebook**
Run locally (i.e., on your laptop)	• **Install R (https://www.r-project.org/)** • **Install RStudio** **Our Choice** **(https://rstudio.com/products/rstudio/download/)**	• Install R (https://www.r-project.org/) • Install RStudio or Jupyter Notebook (https://jupyter.org/)
Run in the cloud	• Option 1: RStudio Cloud (https://posit.cloud/) • Option 2: UofT JupyterHub RStudio (https://jupyter.utoronto.ca/hub/login)	• **Option 1: Google Colab** **Our Choice** **(https://colab.research.google.com/)** • Option 2: UofT JupyterHub Notebook (https://jupyter.utoronto.ca/hub/login)

What’s RStudio?



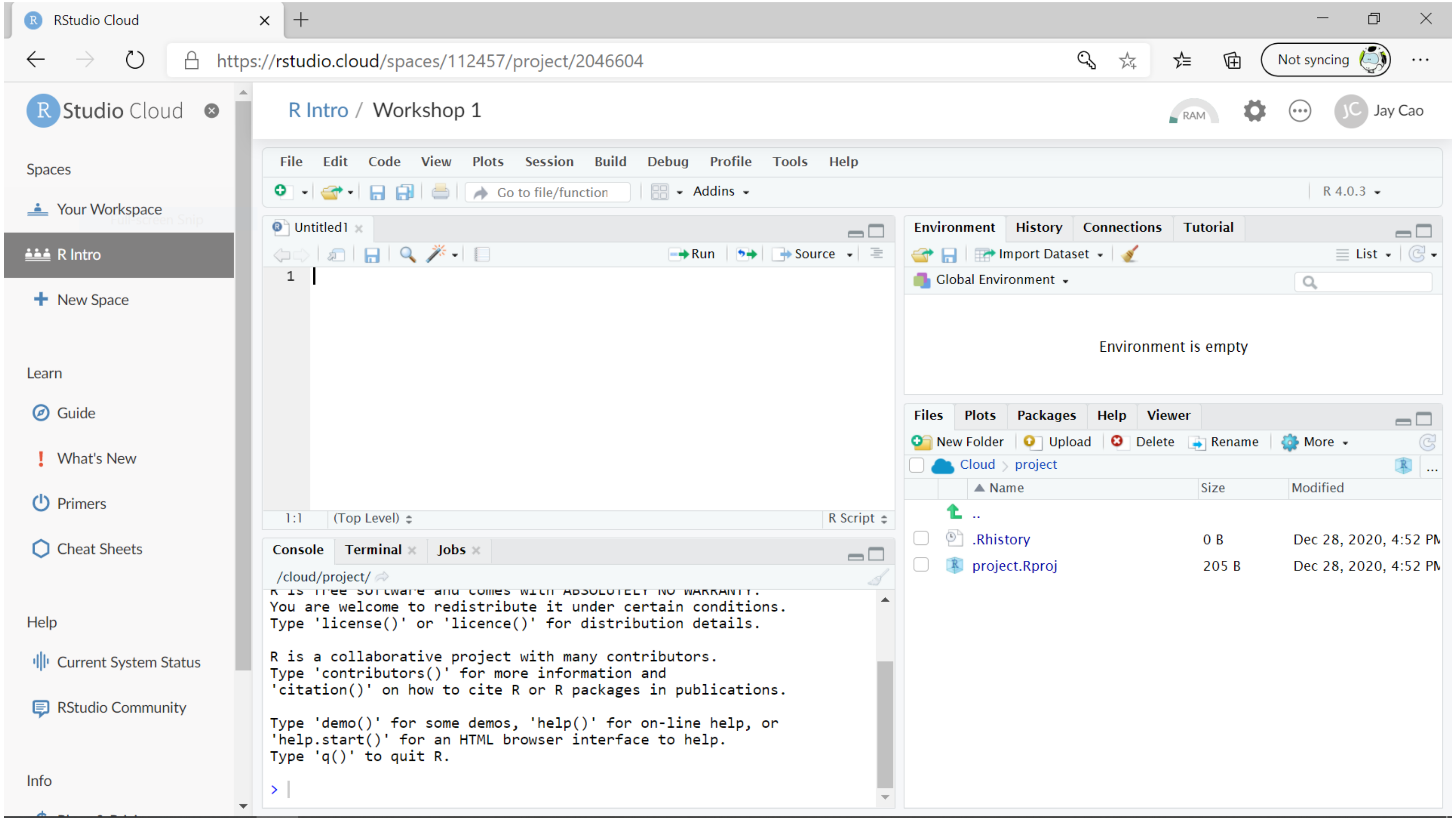
```
library(dplyr)
library(DiagrammeR)

raw <- tribble(
  ~id, ~in_node, ~out_node, ~in_time, ~out_time,
  #--|--|--|
  1, 1, 2, 1, 3,
  1, 2, 3, 3, 5,
  2, 1, 2, 2, 3,
  2, 2, 4, 3, 6
)

node_tb_tp <- raw %>%
  distinct(in_node) %>%
  rename(node_id = in_node)

node_tb <- raw %>%
  distinct(out_node) %>%
  rename(node_id = out_node) %>%
  union(node_tb_tp) %>%
  arrange(node_id)

edge_tb <- raw %>%
  distinct(in_node, out_node) %>%
  rename(from = in_node, to = out_node)

g <- create_graph() %>%
```

Data

edge_tb	3 obs. of 2 variables
g	List of 12
node_tb	4 obs. of 1 variable
node_tb_tp	2 obs. of 1 variable
raw	4 obs. of 5 variables

```
~/OneDrive/rotman/work/mdl/seelab/graph/
+   arrange(node_id)
>
> edge_tb <- raw %>%
+   distinct(in_node, out_node) %>%
+   rename(from = in_node, to = out_node)
>
> g <- create_graph() %>%
+   add_nodes_from_table(table = node_tb) %>%
+   add_edges_from_table(
+     table = edge_tb,
+     from_col = from,
+     to_col = to,
+     from_to_map = node_id)
>
> g %>% render_graph()
>
```

RStudio Cloud



https://rstudio.cloud/spaces/112457/project/2046604

R Intro / Workshop 1

/cloud/project/

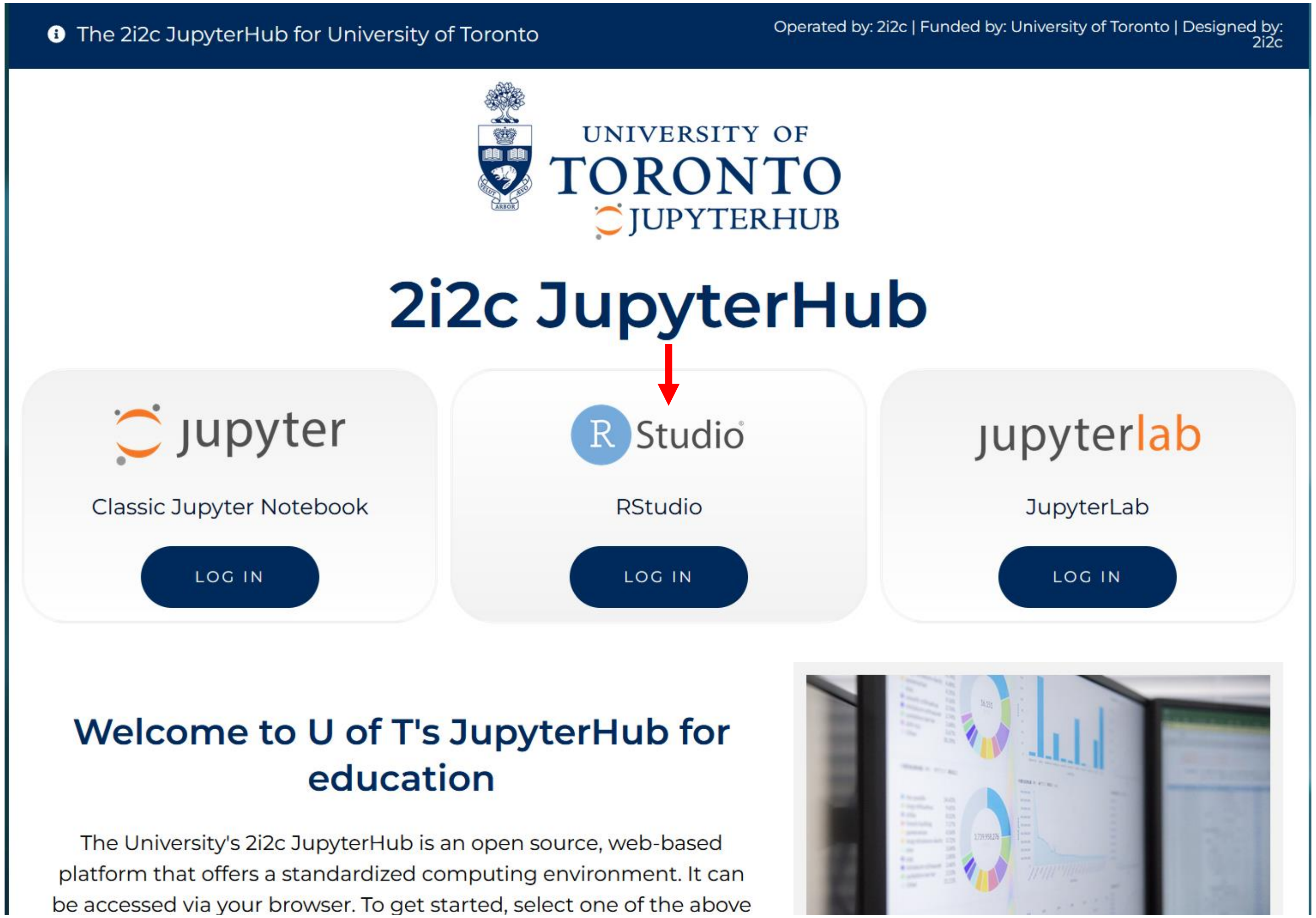
```
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

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 is empty

Cloud > project

Name	Size	Modified
..		
.Rhistory	0 B	Dec 28, 2020, 4:52 PM
project.Rproj	205 B	Dec 28, 2020, 4:52 PM

RStudio at UofT Jupyterhub



The 2i2c JupyterHub for University of Toronto

Operated by: 2i2c | Funded by: University of Toronto | Designed by: 2i2c

UNIVERSITY OF TORONTO JUPYTERHUB

2i2c JupyterHub

Classic Jupyter Notebook

LOG IN

RStudio

LOG IN

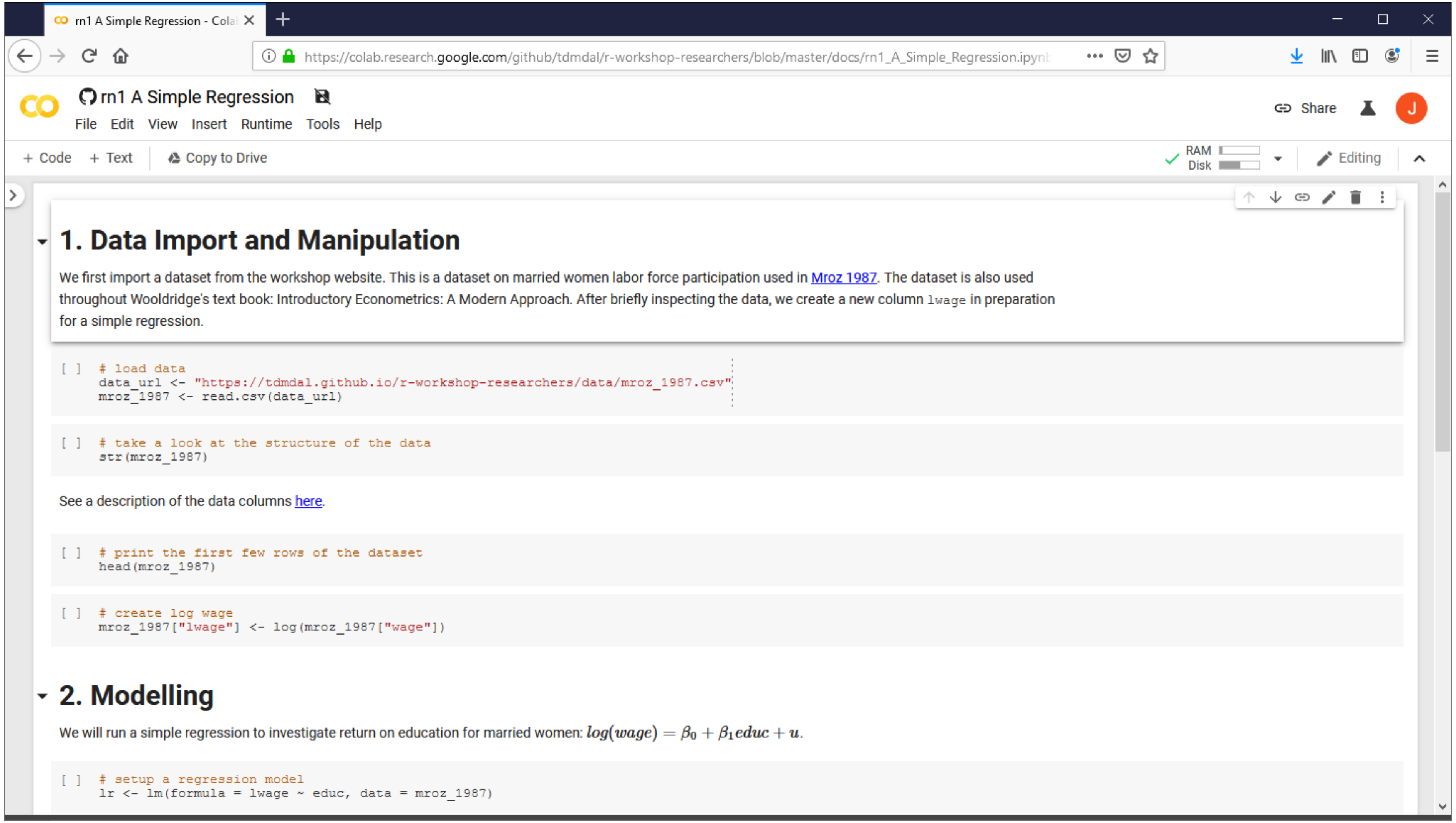
JupyterLab

LOG IN

Welcome to U of T's JupyterHub for education

The University's 2i2c JupyterHub is an open source, web-based platform that offers a standardized computing environment. It can be accessed via your browser. To get started, select one of the above

R Notebook in Google Colab


rn1 A Simple Regression - Cola
https://colab.research.google.com/github/tdmdal/r-workshop-researchers/blob/master/docs/rn1_A_Simple_Regression.ipynb
rn1 A Simple Regression
File
Edit
View
Insert
Runtime
Tools
Help
Share
J
+ Code
+ Text
Copy to Drive
RAM
Disk
Editing

1. Data Import and Manipulation

We first import a dataset from the workshop website. This is a dataset on married women labor force participation used in Mroz 1987. The dataset is also used throughout Wooldridge's text book: Introductory Econometrics: A Modern Approach. After briefly inspecting the data, we create a new column `lwage` in preparation for a simple regression.

```
[ ]  # load data
     data_url <- "https://tdmdal.github.io/r-workshop-researchers/data/mroz_1987.csv"
     mroz_1987 <- read.csv(data_url)
```

```
[ ]  # take a look at the structure of the data
     str(mroz_1987)
```

See a description of the data columns here.

```
[ ]  # print the first few rows of the dataset
     head(mroz_1987)
```

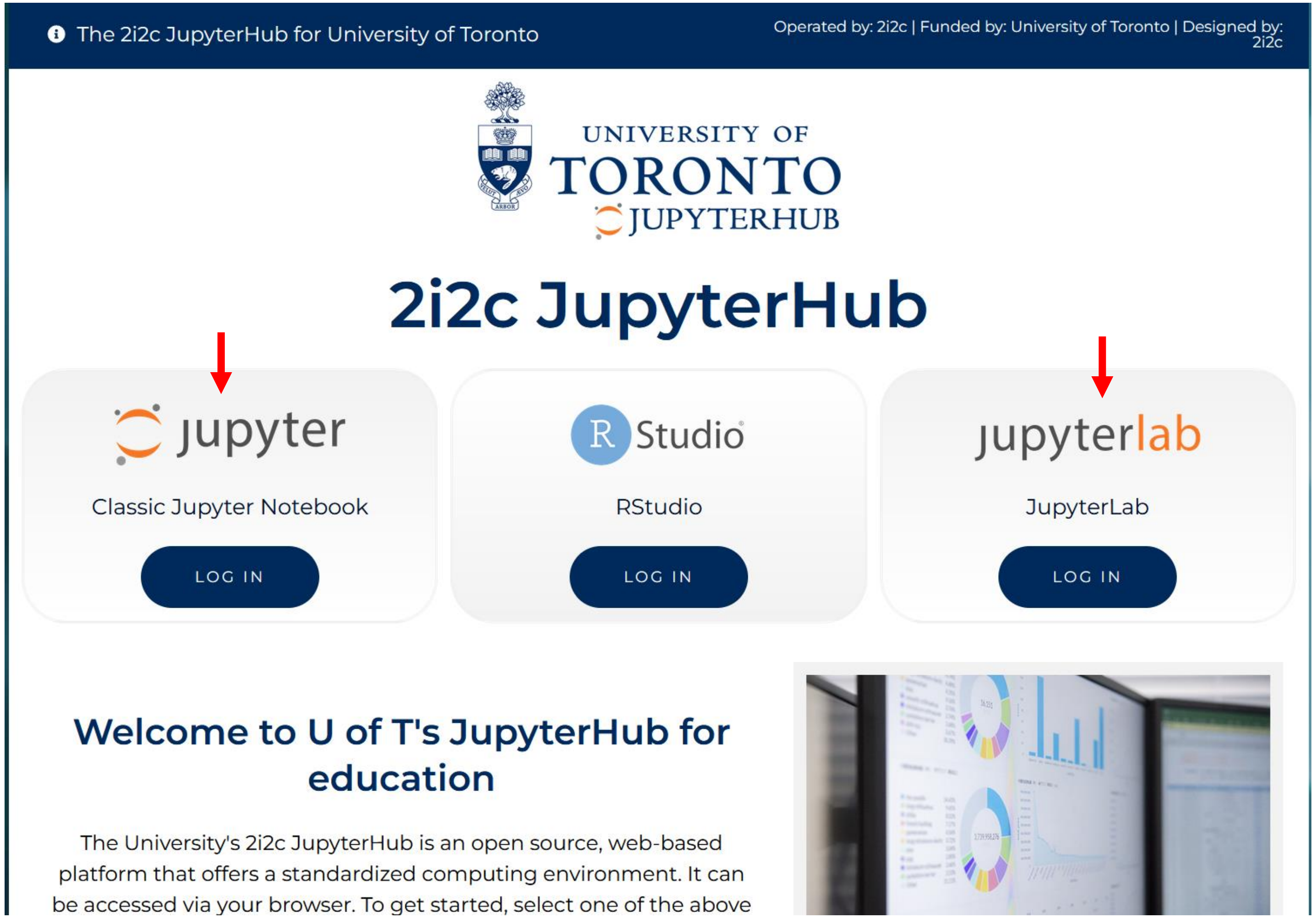
```
[ ]  # create log wage
     mroz_1987["lwage"] <- log(mroz_1987["wage"])
```

2. Modelling

We will run a simple regression to investigate return on education for married women: $log(wage) = \beta_0 + \beta_1 educ + u$.

```
[ ]  # setup a regression model
     lr <- lm(formula = lwage ~ educ, data = mroz_1987)
```

R Notebook at UofT Jupyterhub



The 2i2c JupyterHub for University of Toronto

Operated by: 2i2c | Funded by: University of Toronto | Designed by: 2i2c

2i2c JupyterHub

Welcome to U of T's JupyterHub for education

The University's 2i2c JupyterHub is an open source, web-based platform that offers a standardized computing environment. It can be accessed via your browser. To get started, select one of the above

Plan for Part 1

- Intro
  - What is R and what can R do?
  - Setup R
  - Motivation examples

- Overview of R programing and Data Science
  - Basics of R programming
  - Data science with R

- Learning Resources and Road Map

A Few Examples

- Analyze portfolio performance
- Perform simple sentiment analysis on earning call transcripts
  - Sentiment dictionary approach
  - Language model approach
- Recognize handwritten digits - an example of deep learning

tidyquant

Business Science

PerformanceAnalytics Package

rvest

TIDYTEXT

torch

A Few Examples: What to Look For

- Focus on analysis workflow (by reading the code comments)
  - Import and manipulate data
  - Model data
  - Report and visualize results

- Don’t focus on R syntax
  - By the end of the workshop, you will be able to understand the code

- Do notice everything is done in a sequential way
  - no conditional branching or looping