

# REPRODUCIBLE RESEARCH IN R

SOUMYA BANERJEE

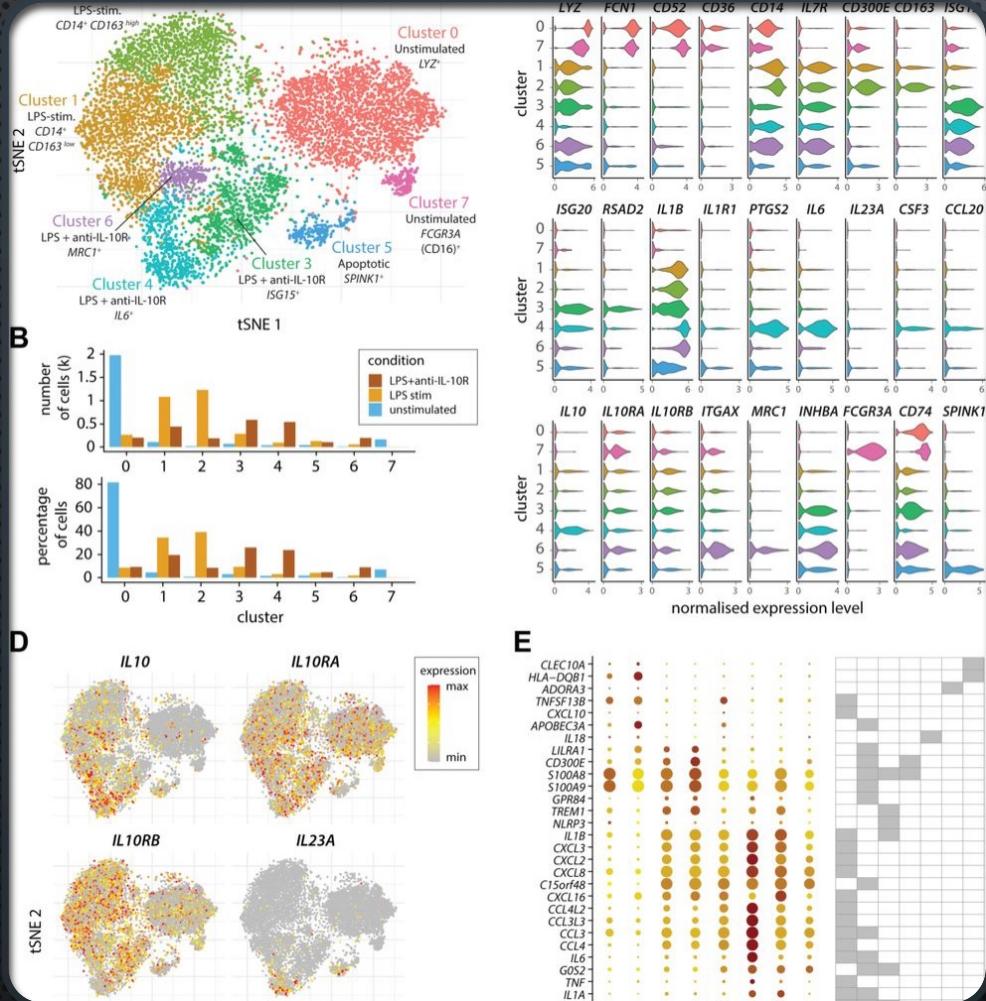
# RATIONALE

- YOUR DATA, YOUR MODEL DECISIONS, PARAMETERS AND YOUR DATA FILTERING DECISIONS WILL KEEP ON CHANGING.
- HOW DO YOU KNOW 6 MONTHS LATER WHAT HAS CHANGED? DOCUMENT YOUR CODE AND YOUR OUTPUT AND YOUR DESIGN DECISIONS ALL IN ONE PLACE.
- REPRODUCIBLE PIPELINE
  - KNOW EXACTLY WHAT CHANGED AND WHEN
  - KNOW HOW TO RERUN THE ANALYSIS AND GET THE (SAME) RESULTS
- THIS IS LIKE YOUR RESEARCH NOTEBOOK

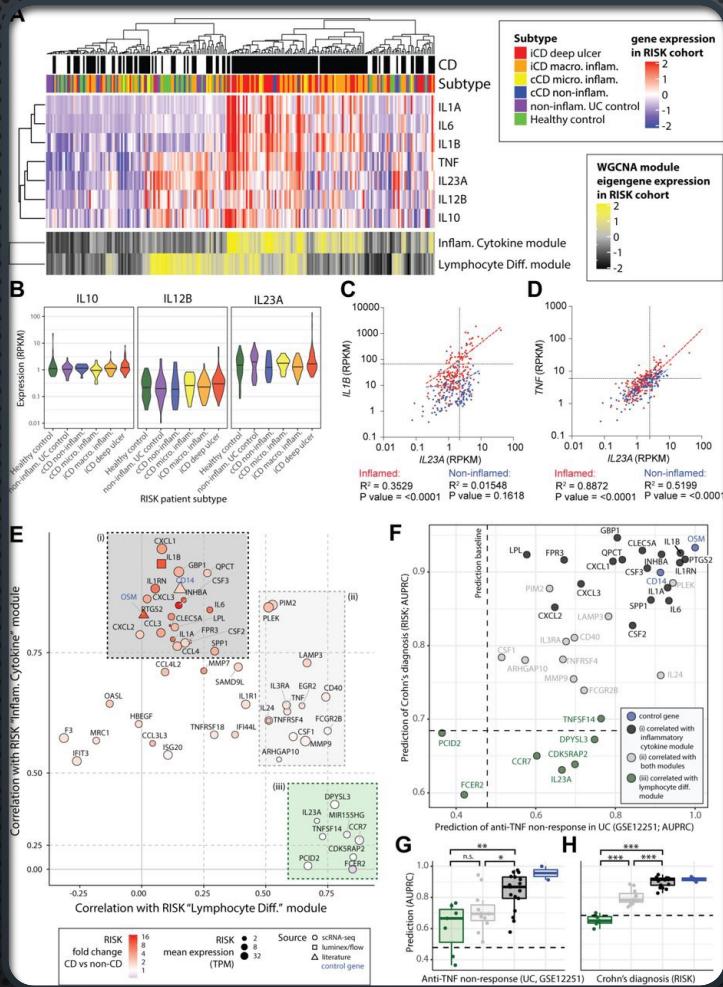
# RATIONALE

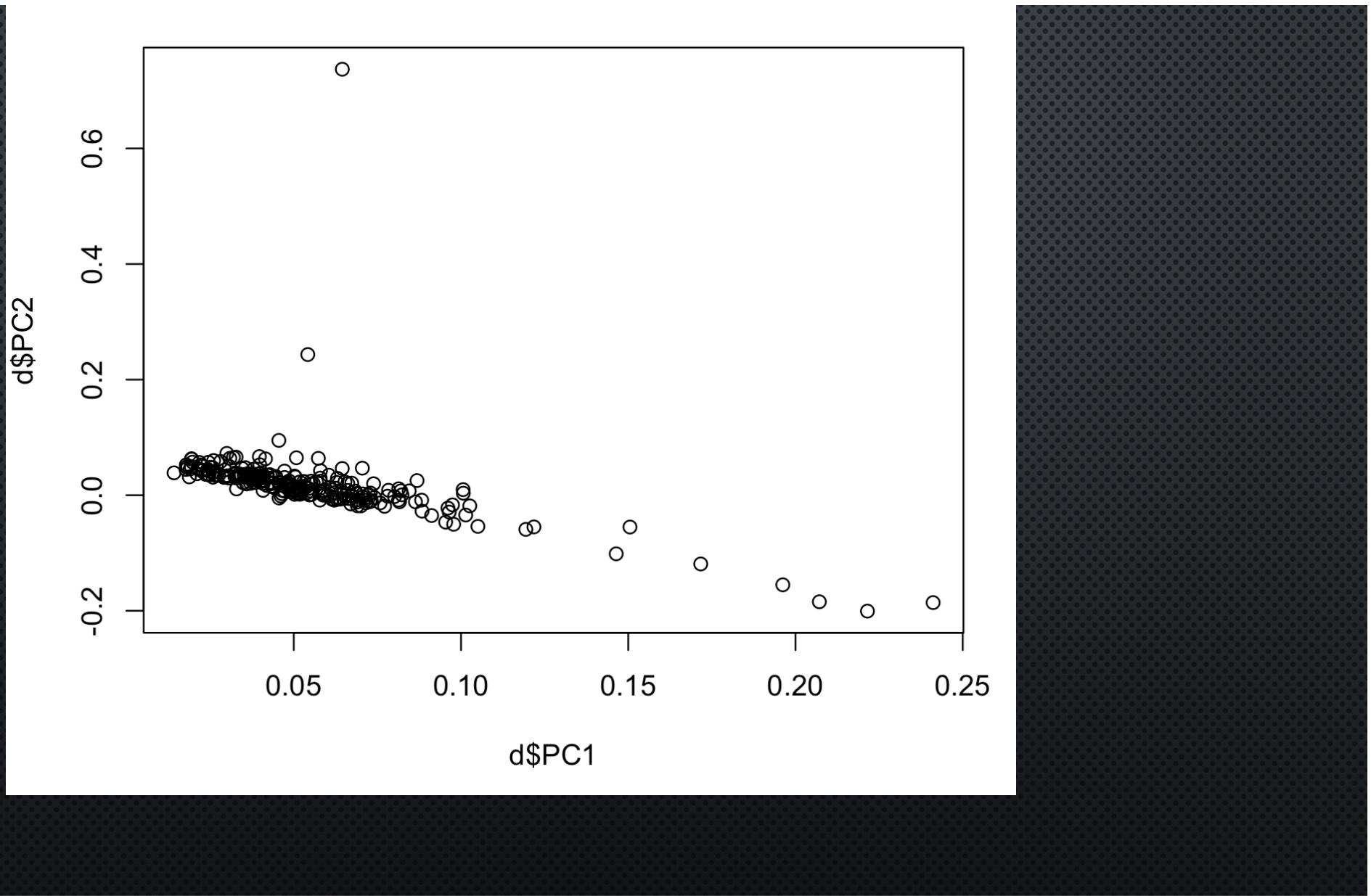
- EXPERIENCES/CASE STUDIES OF USING RMarkdown NOTEBOOKS AND HELPING BIOLOGISTS USE THEM TO ANALYZE THEIR OWN DATA
- WHEN YOU ARE DEEP IN YOUR WORK, IT CAN BE DIFFICULT TO MAKE CODE PRETTY, COMMENT IT AND MAKE IT REPRODUCIBLE.
- BUT YOU WILL REGRET NOT DOING THIS WHEN YOU PARK THE WORK AND 6 MONTHS LATER YOUR COLLABORATORS/REVIEWERS ASK FOR ADDITIONAL ANALYSIS OR CHANGING SOME ASSUMPTION, ETC.
- YOUR CODE SHOULD THEN BE READY (YOU SHOULD BE ABLE TO CLICK A BUTTON AND REPRODUCE THE FIGURES FOR YOUR PAPER).

# CASE STUDY



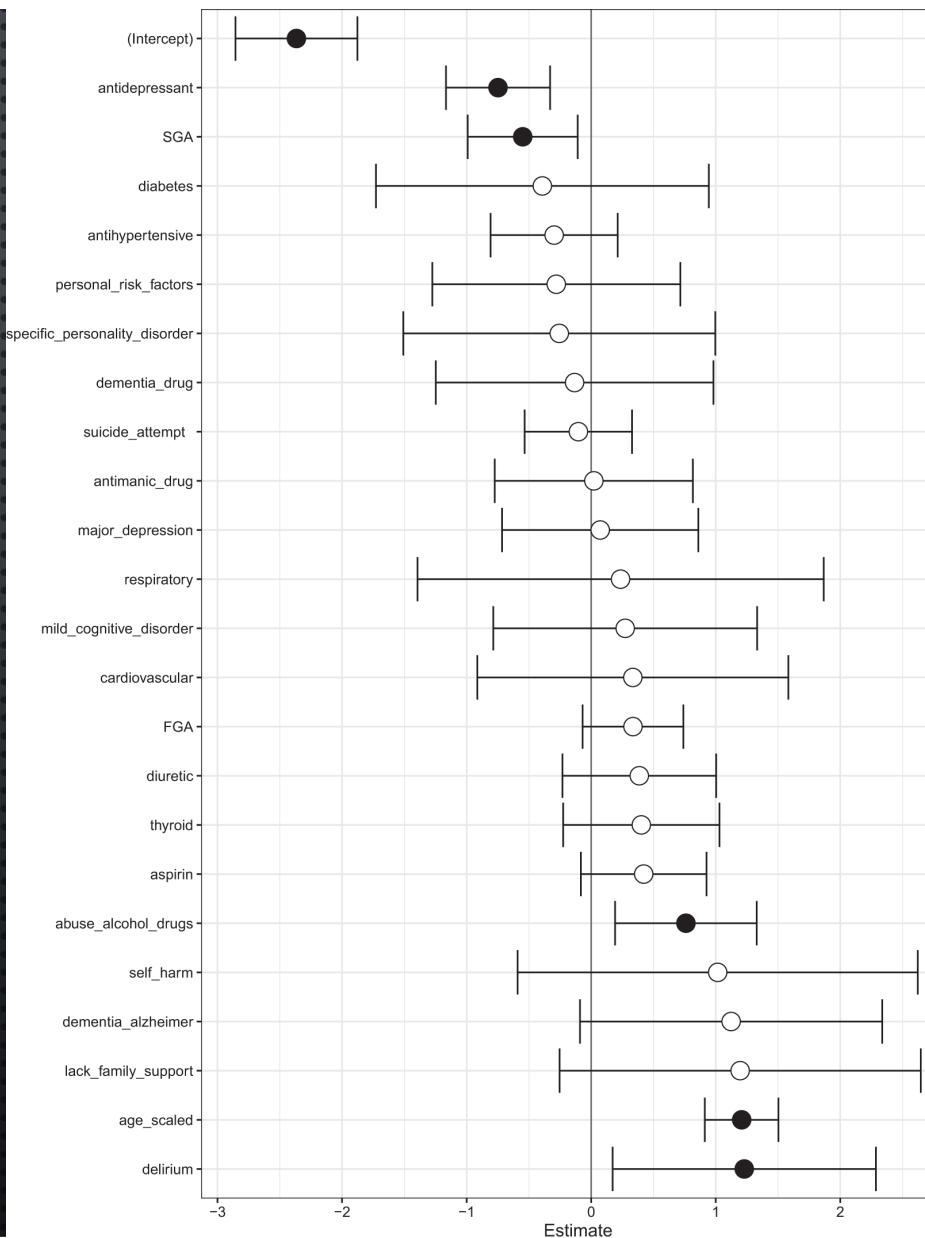
# CASE STUDY

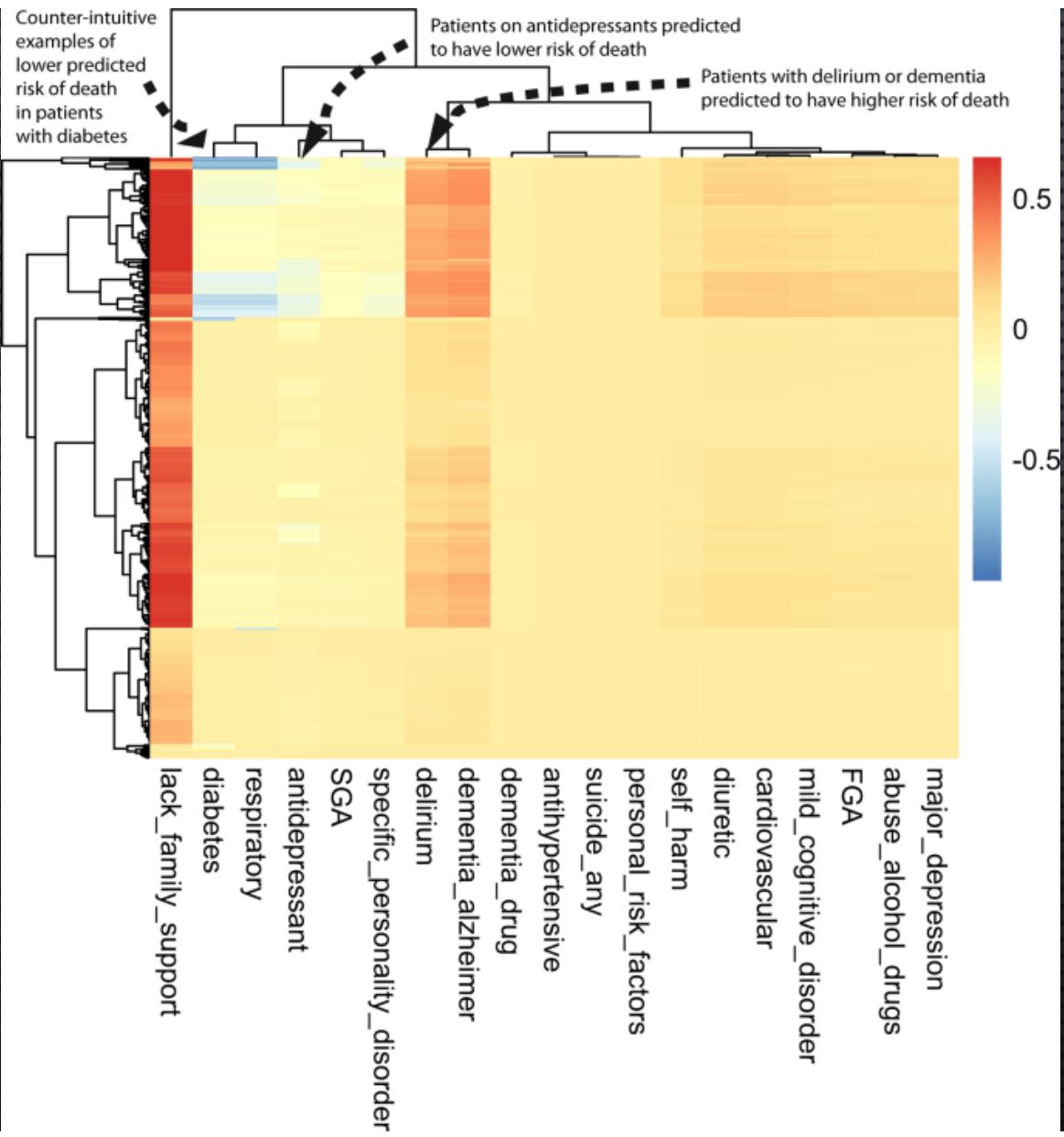


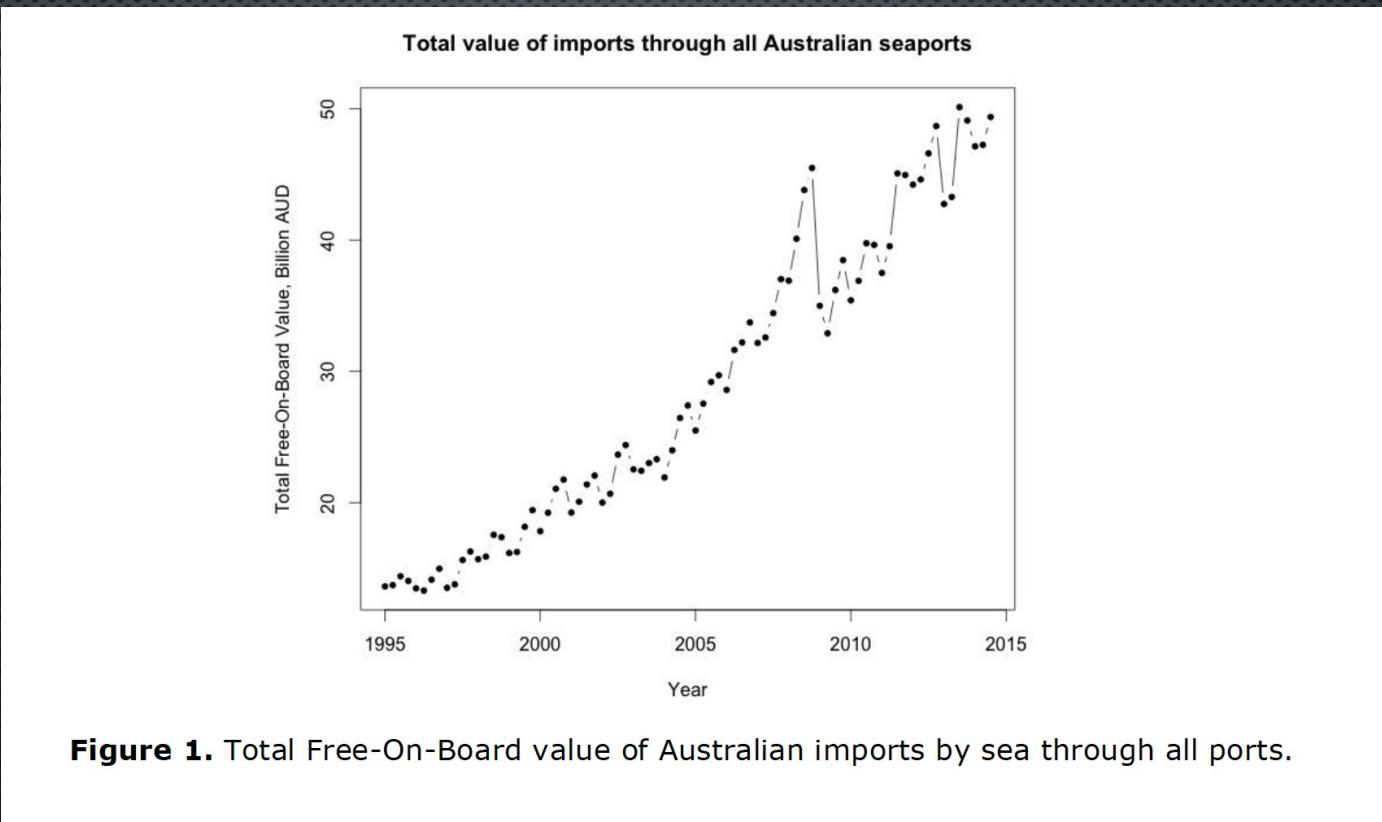


T-TEST

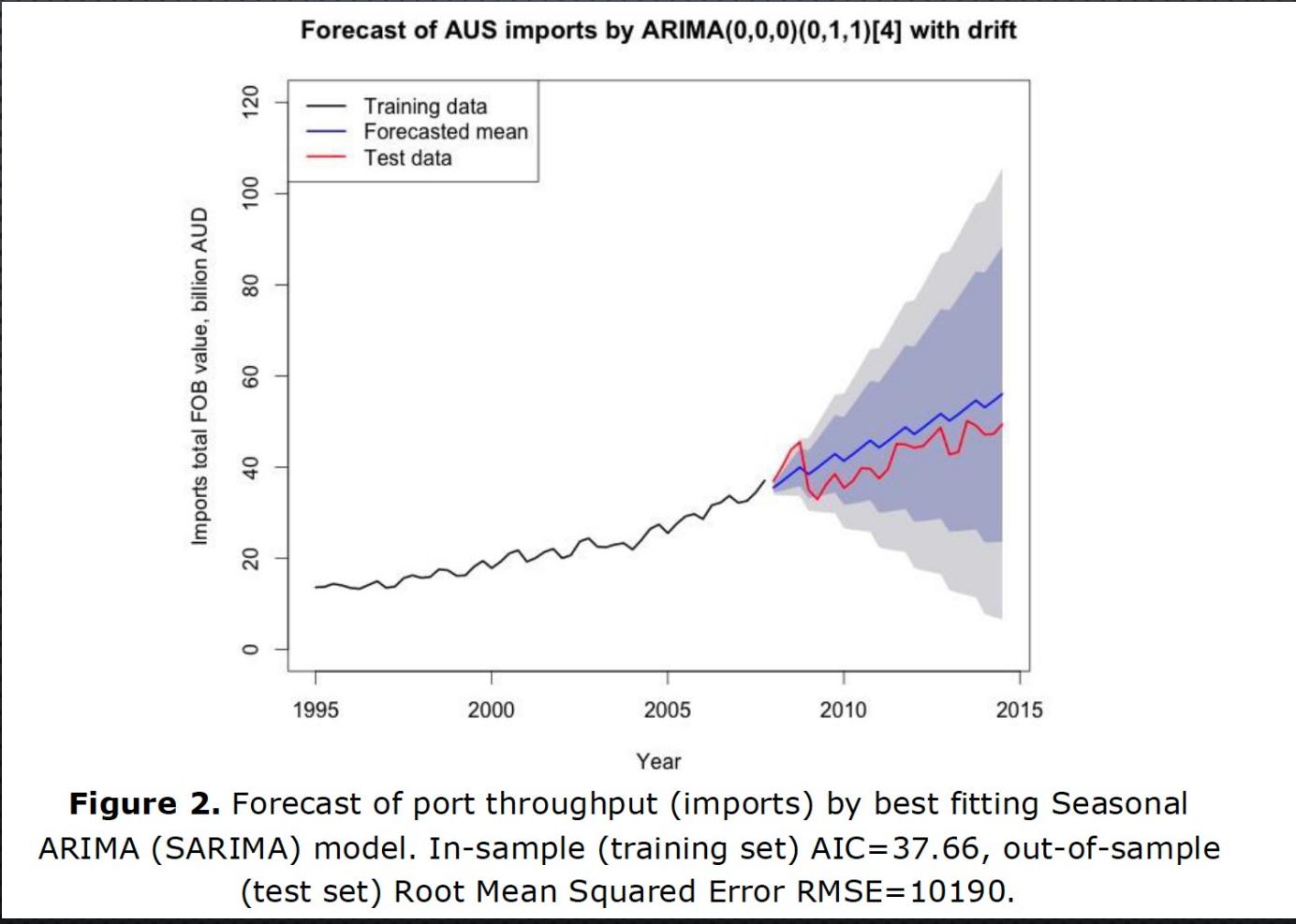
$$\log\left(\frac{p(X)}{1-p(X)}\right)=\beta_0+\beta_1 X.$$

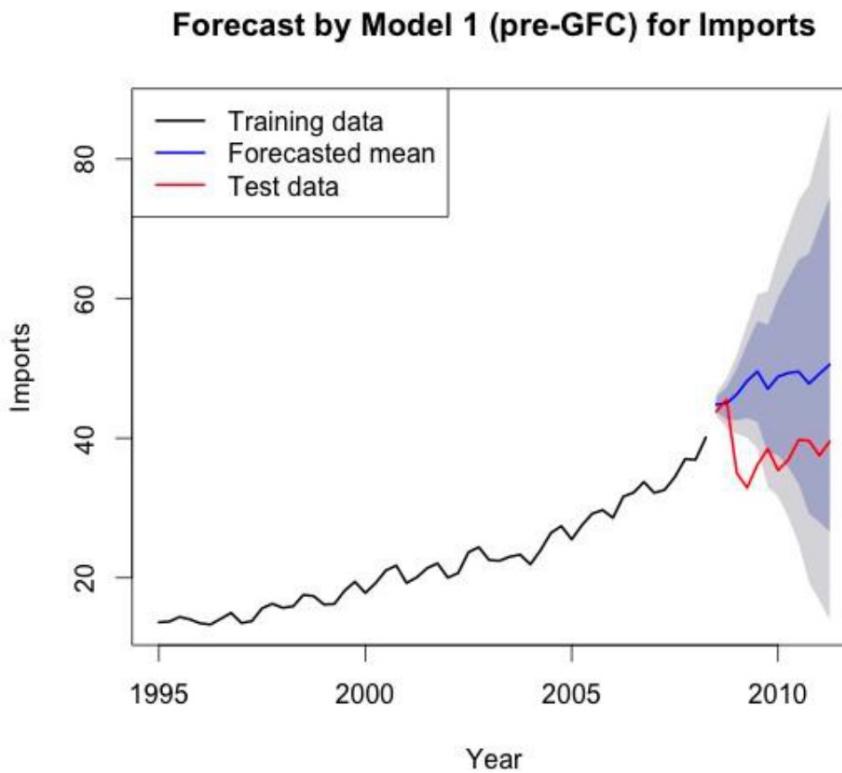






$$y_t = C + \sum_{i=1}^p \phi_i \cdot y_{t-i} - \sum_{i=1}^q \theta_i \cdot \epsilon_{t-i}$$





**Figure 6.** Forecast of port throughput by Model 1 (7D VAR model using GDPs of Australia and of top-5 importers as additional predictors of imports) trained on pre-GFC data from Q1 1995 to Q2 2008. Out-of-sample (test set from Q3 2008 to Q2 2011) Root Mean Squared Error RMSE=10685.

*Vector Auto-Regressive (VAR) model:*

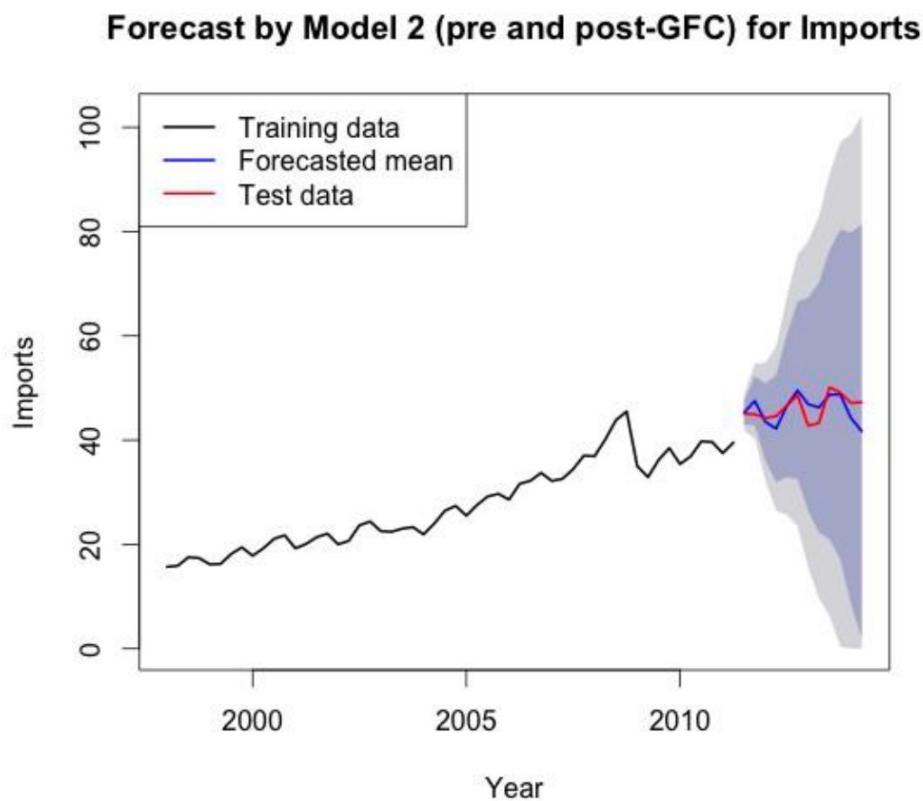
This is a multivariate model, which is capable of modelling the joint dependencies between the throughput and the supporting GDP data, and uses these dependencies to forecast the imports along with the GDP in the future.

A  $p$ -th order VAR( $p$ ) is represented by the following equation:

$$\mathbf{y}_t = \mathbf{c} + \sum_{i=1}^p A_i \cdot \mathbf{y}_{t-i} + \epsilon_t$$

The variables (port throughput and GDP of all countries) are subsumed in the vector  $y_t$ , and  $y_{t-i}$  is the  $i$ th lag of  $y_t$ . The coefficient matrices  $A_i$  are time-invariant and represent a set of model parameters,  $\epsilon$  is a vector of error terms with mean 0 and covariance  $\Sigma$ , and  $c$  is a vector of constant intercept terms. Fitting the VAR model involves estimating the matrix of interactions  $A_i$ , vector  $c$  and the covariance matrix  $\Sigma$  using the training data.

In this work, two different types of VAR model classes are considered: a two dimensional class which incorporates the imports time series, and the GDP of Australia only; and a seven dimensional model class that in addition to the port throughput data and Australian GDP

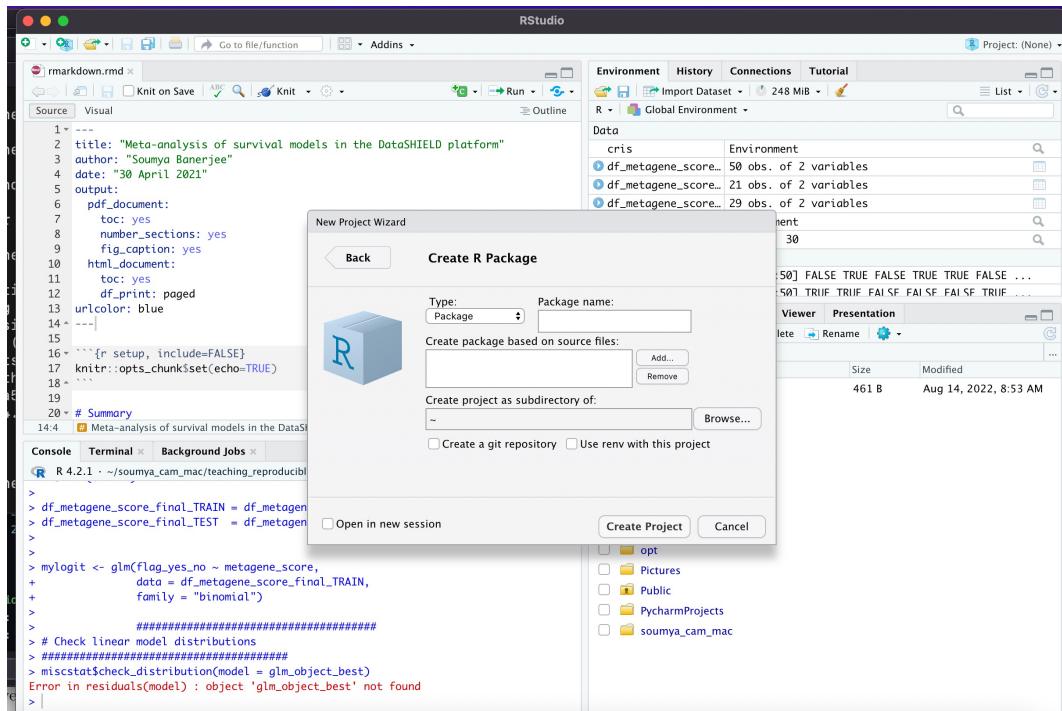


**Figure 7.** Forecast of port throughput by Model 2 (7D VAR model using GDPs of Australia and of top-5 importers as additional predictors of imports) trained on pre- and post-GFC data from Q1 1998 to Q2 2011. Out-of-sample (test set from Q3 2011 to Q2 2014) Root Mean Squared Error RMSE=2594.

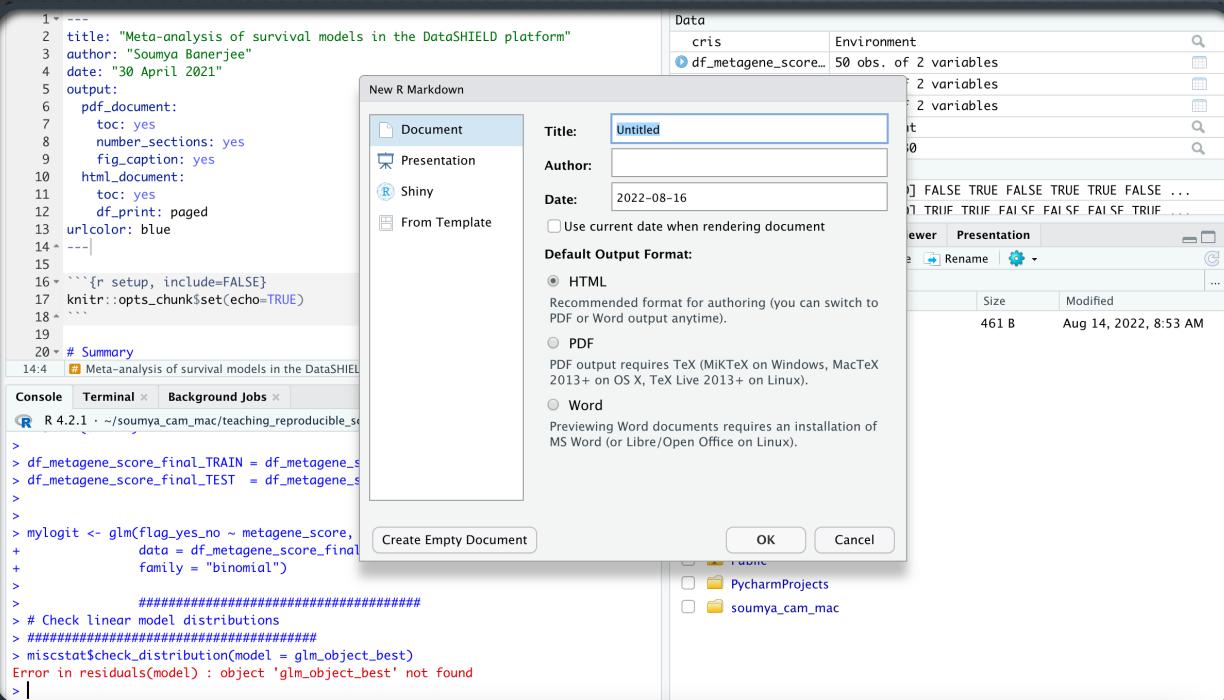
# SHINY APPS FOR RAPID PROTOTYPING AND COMMUNICATION

# SHINY APPS FOR RAPID PROTOTYPING AND COMMUNICATION

# CREATING PACKAGES



# CREATING R MARKDOWN



# DEMO

The screenshot shows the RStudio interface with the following components:

- Left Panel (Code Editor):** An R Markdown file titled "rmarkdown.rmd". The code includes setup code, a header, and examples of R code chunks and plots.
- Middle Panel (Environment):** The "Environment" pane displays the message "Environment is empty".
- Right Panel (File Explorer):** The "Files" pane shows the user's home directory structure.

```
1 ---  
2 title: "Untitled"  
3 output: html_document  
4 date: "2022-08-16"  
5 ---  
6  
7 ```{r setup, include=FALSE}  
8 knitr::opts_chunk$set(echo = TRUE)  
9 ```  
10 ## R Markdown  
11  
12 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>.  
13  
14 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:  
15  
16 ```{r cars}  
17 summary(cars)  
18 ```  
19  
20  
21 ## Including Plots  
22  
23 You can also embed plots, for example:  
24  
25 ```{r pressure, echo=FALSE}  
26 plot(pressure)  
27 ```  
28  
29 Note that the 'echo = FALSE' parameter was added to the code chunk to prevent printing of the R code that generated the plot.  
30
```

# R MARKDOWN BASICS

- ````R { }```
- INCLUDE = FALSE PREVENTS CODE AND RESULTS FROM APPEARING IN THE FINISHED FILE. R MARKDOWN STILL RUNS THE CODE IN THE CHUNK, AND THE RESULTS CAN BE USED BY OTHER CHUNKS.
- ECHO = FALSE PREVENTS CODE, BUT NOT THE RESULTS FROM APPEARING IN THE FINISHED FILE. THIS IS A USEFUL WAY TO EMBED FIGURES.
- MESSAGE = FALSE PREVENTS MESSAGES THAT ARE GENERATED BY CODE FROM APPEARING IN THE FINISHED FILE.
- WARNING = FALSE PREVENTS WARNINGS THAT ARE GENERATED BY CODE FROM APPEARING IN THE FINISHED.
- FIG.CAP = "..." ADDS A CAPTION TO GRAPHICAL RESULTS.

## Syntax

Make a code chunk with three back ticks followed by an r in braces. End the chunk with three back ticks:

```
```{r}
paste("Hello", "World!")
```
```

Place code inline with a single back ticks. The first back tick must be followed by an R, like this `r paste("Hello", "World!")`.

Add chunk options within braces. For example, `echo=FALSE` will prevent source code from being displayed:

```
```{r eval=TRUE, echo=FALSE}
paste("Hello", "World!")
```
```

## Becomes

Make a code chunk with three back ticks followed by an r in braces. End the chunk with three back ticks:

```
paste("Hello", "World!")
```

```
## [1] "Hello World!"
```

Place code inline with a single back ticks. The first back tick must be followed by an R, like this Hello World!.

Add chunk options within braces. For example, `echo=FALSE` will prevent source code from being displayed:

```
## [1] "Hello World!"
```

# DEMO

# RATIONALE

- THE CONCEPTS ARE THE SAME IN ANY PROGRAMMING LANGUAGE (R/PYTHON)
- BOTTOMLINE: WE ARE ALL BUSY AND WE WOULD ALL RATHER PUBLISH PAPERS, BUT IN THE LONG TERM THESE BEST PRACTICES WILL MAKE US MORE PRODUCTIVE
- THIS IS LIKE PROTOCOLS (USED IN EXPERIMENTAL BIOLOGY) FOR COMPUTER SCIENTISTS. ALSO LIKE A LAB NOTEBOOK BUT FOR COMPUTATIONAL PEOPLE.

# CONCEPTS ARE LANGUAGE AGNOSTIC

```
# This is a test
```

```
Testing pandoc
```

```
``` code
import GPy
```
```

```
``` code
print("Cat")
```
```

# CONCEPTS ARE LANGUAGE AGNOSTIC

- [HTTPS://GITHUB.COM/NEELSOUMYA/TEACHING\\_REPRODUCIBLE\\_SCIENCE\\_R/BLOB/MAIN/TST.MD](https://github.com/neelsoumya/teaching_reproducible_science_R/blob/main/tst.md)
- *PANDOC TST.MD -O TEST.IPYNB*

# DEMO

- SEE THE LINK BELOW FOR MORE DETAILS
- [HTTPS://WWW.RSTUDIO.COM/WP-CONTENT/UPLOADS/2015/03/RMarkdown-REFERENCE.PDF](https://www.rstudio.com/wp-content/uploads/2015/03/rmarkdown-reference.pdf)
- NOW HEAD OVER TO THE FILE NAMED RMarkdown.RMD
- [HTTPS://GITHUB.COM/NEELSOUMYA/TEACHING\\_REPRODUCIBLE\\_SCIENCE\\_R/BLOB/MAIN/RMARKDOWN.RMD](https://github.com/NeelSoumya/Teaching_Reproducible_Science_R/blob/main/rmarkdown.Rmd)
- RUNNING THIS WILL CREATE A REPORT LIKE THE FOLLOWING:
- [HTTPS://GITHUB.COM/NEELSOUMYA/TEACHING\\_REPRODUCIBLE\\_SCIENCE\\_R/BLOB/MAIN/RMARKDOWN.PDF](https://github.com/NeelSoumya/Teaching_Reproducible_Science_R/blob/main/rmarkdown.pdf)

# DEMO WITH RMARKDOWN FROM REAL-WORLD PROJECT

# EXERCISES WITH SYNTHETIC DATA

# GRAPHICAL USER INTERFACES

- YOU CAN ALSO EASILY CREATE GRAPHICAL USER INTERFACES HERE IS A DEMO:  
[HTTPS://SB2333MEDSCHL.SHINYAPPS.IO/SHINYAPP](https://sb2333medschl.shinyapps.io/shinyapp)
- CODE APP.R:  
[HTTPS://GITHUB.COM/NEELSOUMYA/TEACHING\\_REPRODUCIBLE\\_SCIENCE\\_R/BLOB/MAIN/SHINYAPP/APP.R](https://github.com/neelsoumya/TEACHING_REPRODUCIBLE_SCIENCE_R/blob/main/shinyapp/app.R)
- DEMO

# RESOURCES

- [HTTPS://GITHUB.COM/NEELSOUMYA/TEACHING\\_REPRODUCIBLE\\_SCIENCE\\_R/](https://github.com/neelsoumya/TEACHING_REPRODUCIBLE_SCIENCE_R/)
- CODE, RESOURCES AND TEMPLATES
- EXERCISES WITH SYNTHETIC DATA
- INSTALLATION INSTRUCTIONS