# REPRODUCIBLE RESEARCH IN R

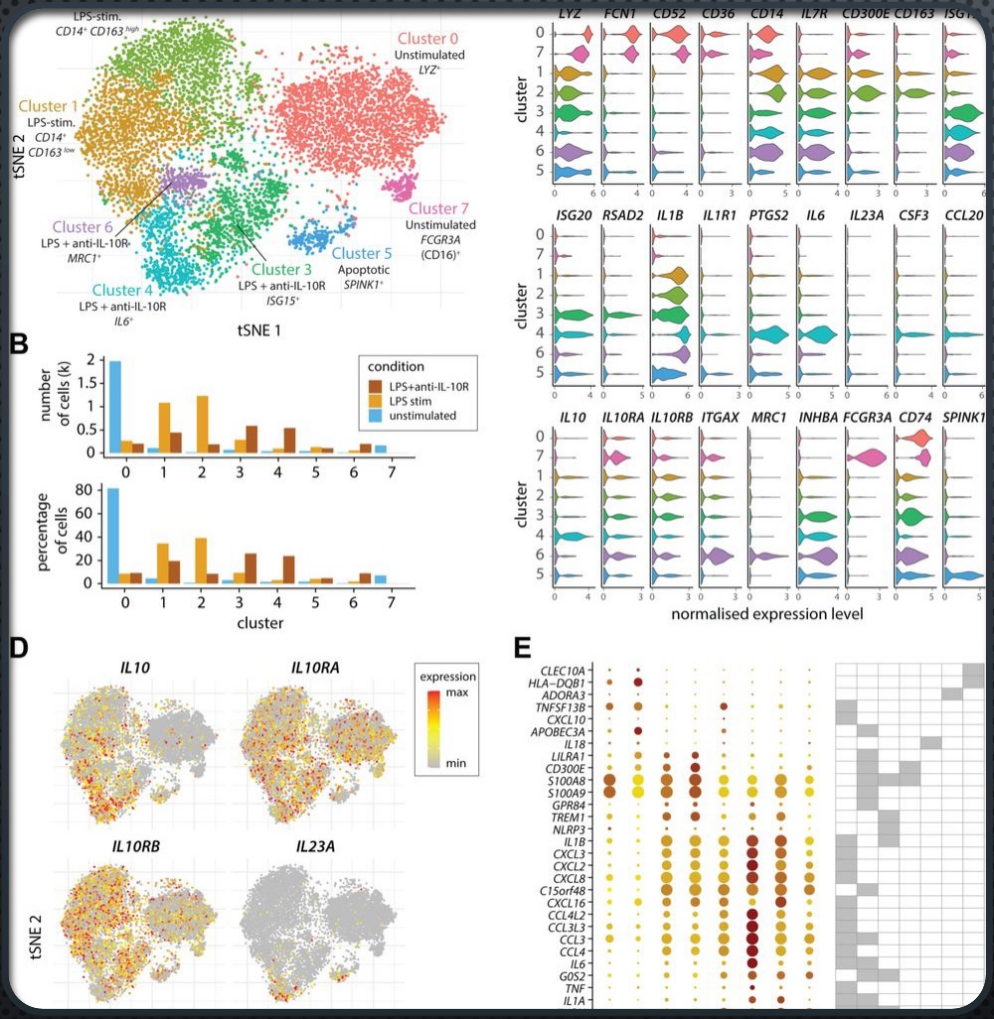
SOUMYA BANERJEE

# Rationale

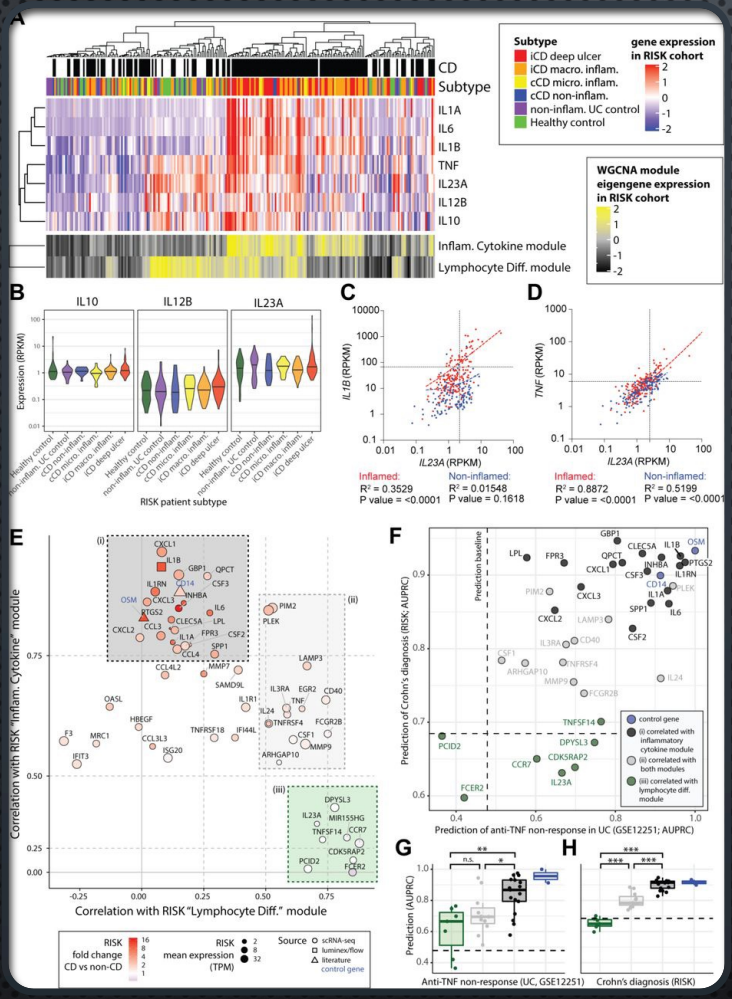
- Your data, your model decisions, parameters and your data filtering decisions will keep on changing.
- How do you know 6 months later what has changed? Document your code and your output and your design decisions all in one place.
- Reproducible pipeline
  - Know exactly what changed and when
  - Know how to rerun the analysis and get the (same) results
- This is like your research notebook

# RATIONALE

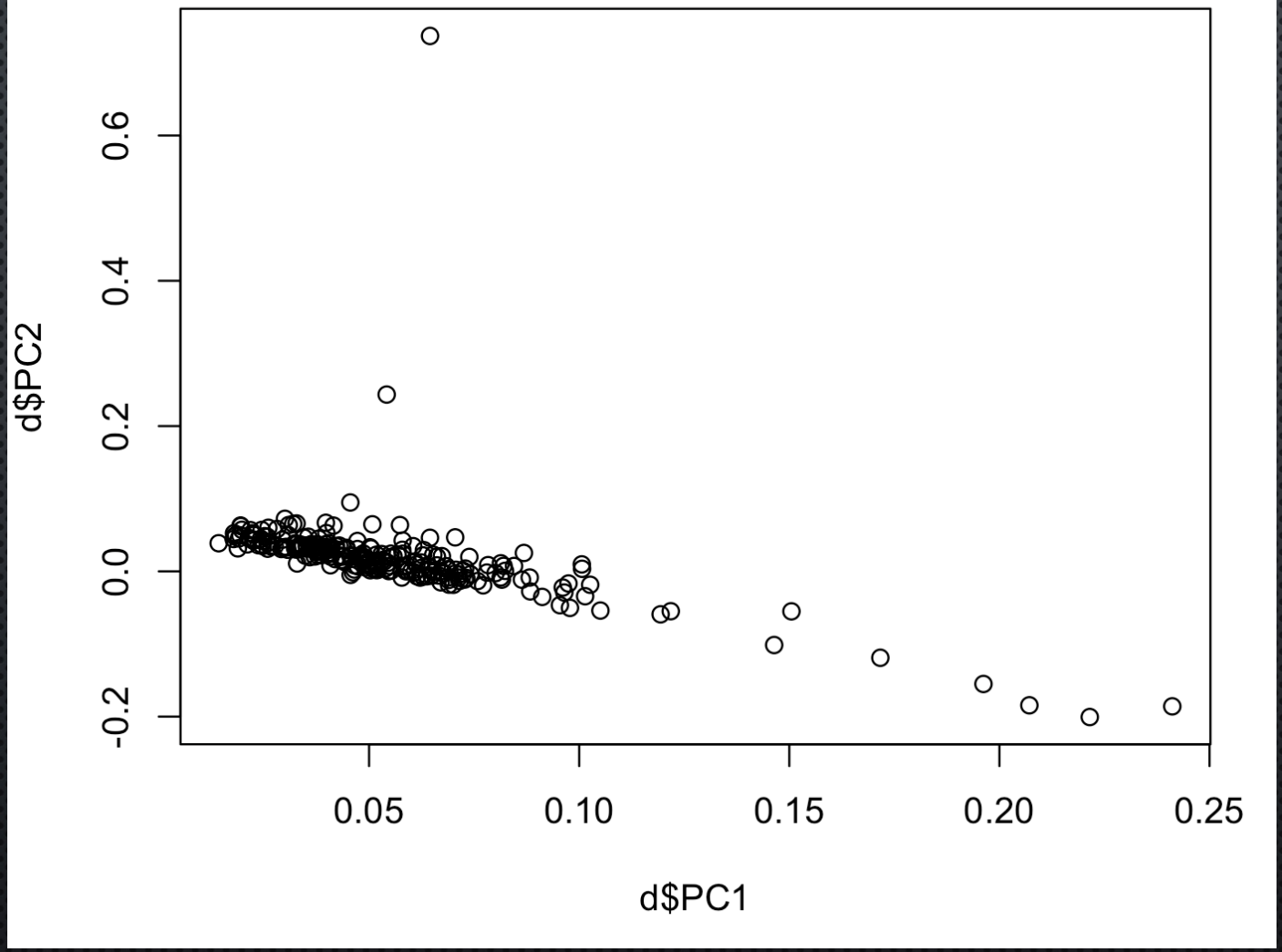
- experiences/case studies of using Rmarkdown notebooks and helping biologists use them to analyze their own data
- When you are deep in your work, it can be difficult to make code pretty, comment it and make it reproducible.
- But you will regret not doing this when you park the work and 6 months later your colloborators/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).


Cluster 0
Unstimulated
LYZ+
Cluster 1
LPS-stim.
CD14+
CD163low
Cluster 6
LPS + anti-IL-10R
MRC1+
Cluster 4
LPS + anti-IL-10R
IL6+
Cluster 3
LPS + anti-IL-10R
ISG15+
Cluster 5
Apoptotic
SPINK1+
Cluster 7
Unstimulated
FCGR3A
(CD16)+
tSNE 1
tSNE 2
B
number of cells (k)
percentage of cells
cluster
condition
LPS+anti-IL-10R
LPS stim
unstimulated
LY2 FCN1 CD52 CD36 CD14 IL7R CD300E CD163
ISG20 RSAD2 IL1B IL1R1 PTGS2 IL6 IL23A CSF3 CCL20
IL10 IL10RA IL10RB ITGAX MRC1 INHBA FCGR3A CD74 SPINK1
normalised expression level
D
IL10
IL10RA
IL10RB
IL23A
expression
max
min
E
CLEC10A
HLA-DQB1
ADORA3
TNFSF13B
CXCL10
APOBEC3A
IL18
LILRA1
CD300E
S100A8
S100A9
GPR84
TREM1
NLRP3
IL1B
CXCL3
CXCL2
CXCL8
C15orf48
CXCL16
CCL4L2
CCL3L3
CCL3
CCL4
IL6
G0S2
TNF
IL1A

# CASE STUDY


A
CD
Subtype
IL1A
IL6
IL1B
TNF
IL23A
IL12B
IL10
Inflam. Cytokine module
Lymphocyte Diff. module
Subtype
iCD deep ulcer
iCD macro. inflam.
cCD micro. inflam.
cCD non-inflam.
non-inflam. UC control
Healthy control
gene expression in RISK cohort
WGCNA module eigengene expression in RISK cohort
B
IL10
IL12B
IL23A
Expression (RPKM)
RISK patient subtype
C
IL1B (RPKM)
IL23A (RPKM)
Inflamed:
R² = 0.3529
P value = <0.0001
Non-inflamed:
R² = 0.01548
P value = 0.1618
D
TNF (RPKM)
IL23A (RPKM)
Inflamed:
R² = 0.8872
P value = <0.0001
Non-inflamed:
R² = 0.5199
P value = <0.0001
E
Correlation with RISK "Inflam. Cytokine" module
Correlation with RISK "Lymphocyte Diff." module
RISK fold change CD vs non-CD
RISK mean expression (TPM)
Source
scRNA-seq
luminex/flow
literature
control gene
F
Prediction baseline
Prediction of Crohn's diagnosis (RISK; AUPRC)
Prediction of anti-TNF non-response in UC (GSE12251; AUPRC)
control gene
(i) correlated with inflammatory cytokine module
(ii) correlated with both modules
(iii) correlated with lymphocyte diff. module
G
Prediction (AUPRC)
Anti-TNF non-response (UC, GSE12251)
H
Crohn's diagnosis (RISK)

# CASE STUDY


d$PC2
0.6
0.4
0.2
0.0
-0.2
0.05
0.10
0.15
0.20
0.25
d$PC1

# T-TEST

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

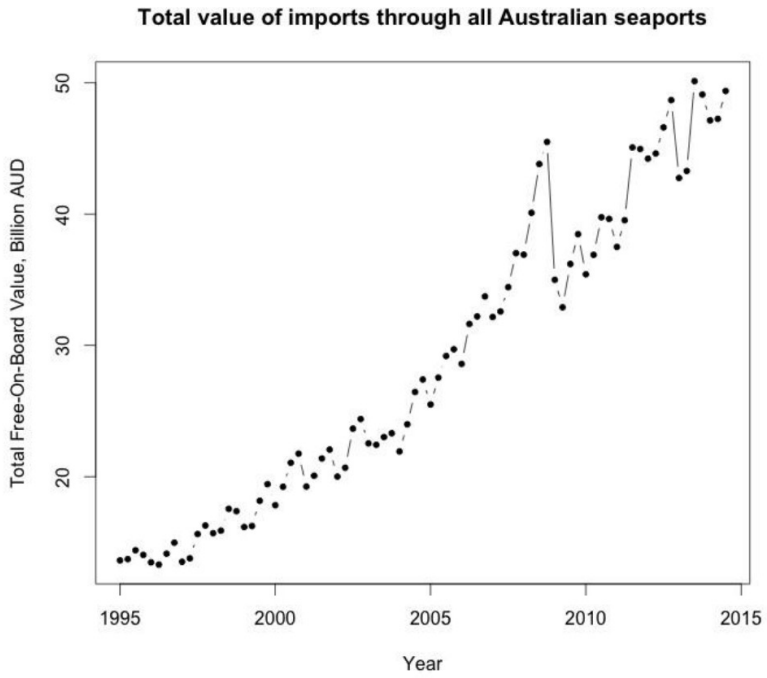



**Figure 1.** Total Free-On-Board value of Australian imports by sea through all ports.

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

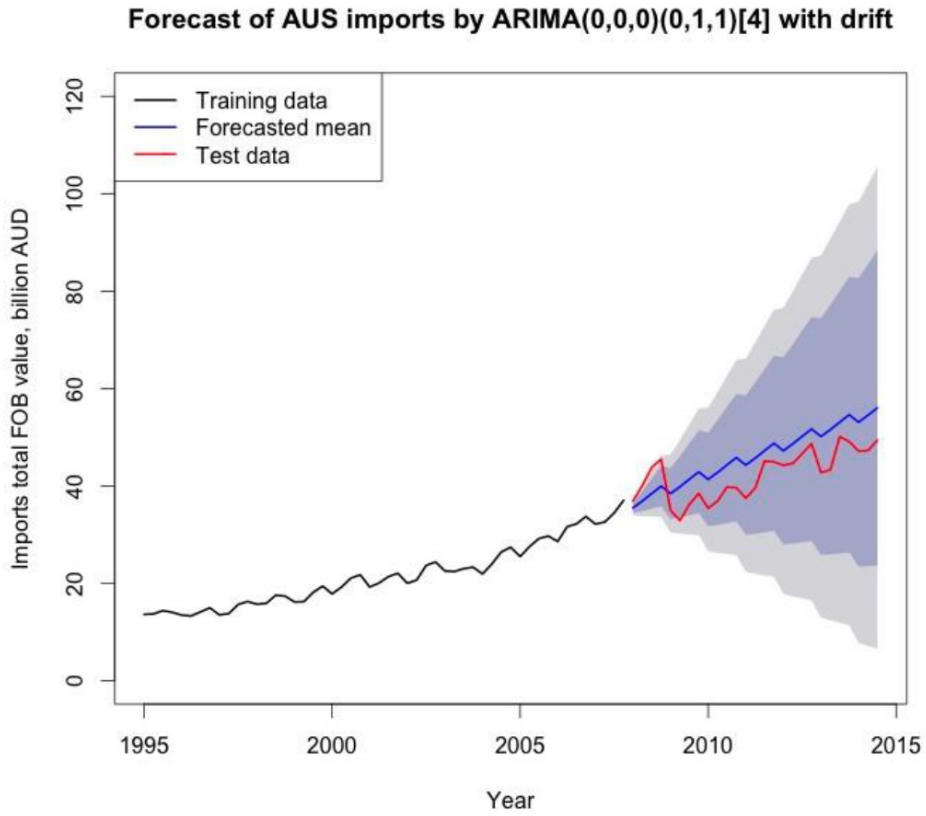



**Figure 2.** Forecast of port throughput (imports) by best fitting Seasonal ARIMA (SARIMA) model. In-sample (training set) AIC=37.66, out-of-sample (test set) Root Mean Squared Error RMSE=10190.

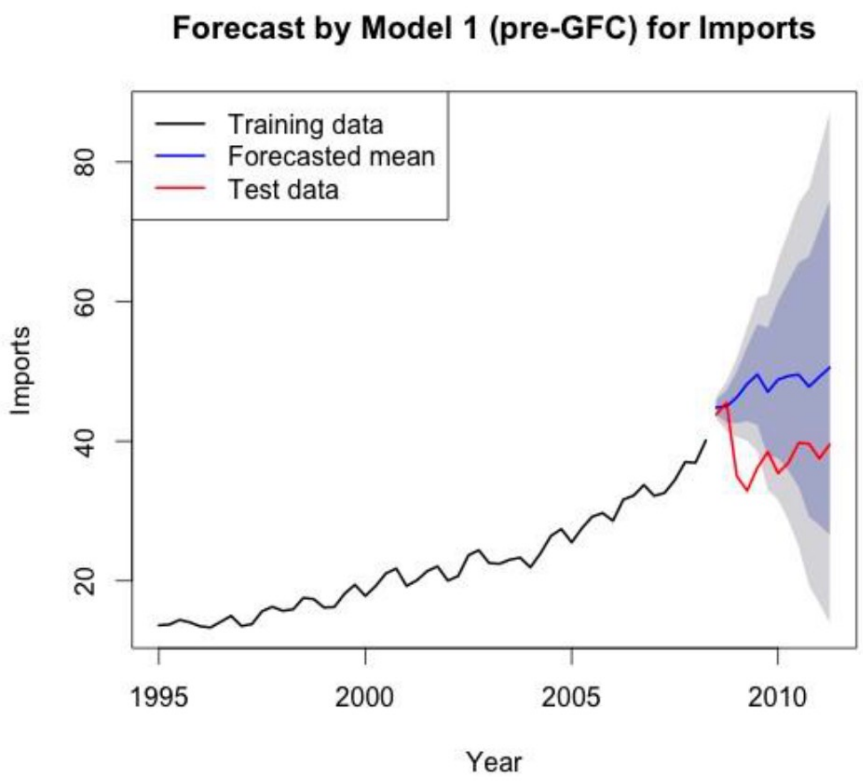



**Figure 6**. Forecast of port throughput by Model 1 (7D VAR model using GDPs of Australia and of top-5 importers as additional predictors of imports) trained on pre-GFC data from Q1 1995 to Q2 2008. Out-of-sample (test set from Q3 2008 to Q2 2011) Root Mean Squared Error RMSE=10685.

*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 $y_!$, and $y_{!!!}$ is the $i$th lag of $y_!$. The coefficient matrices $A_!$ 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

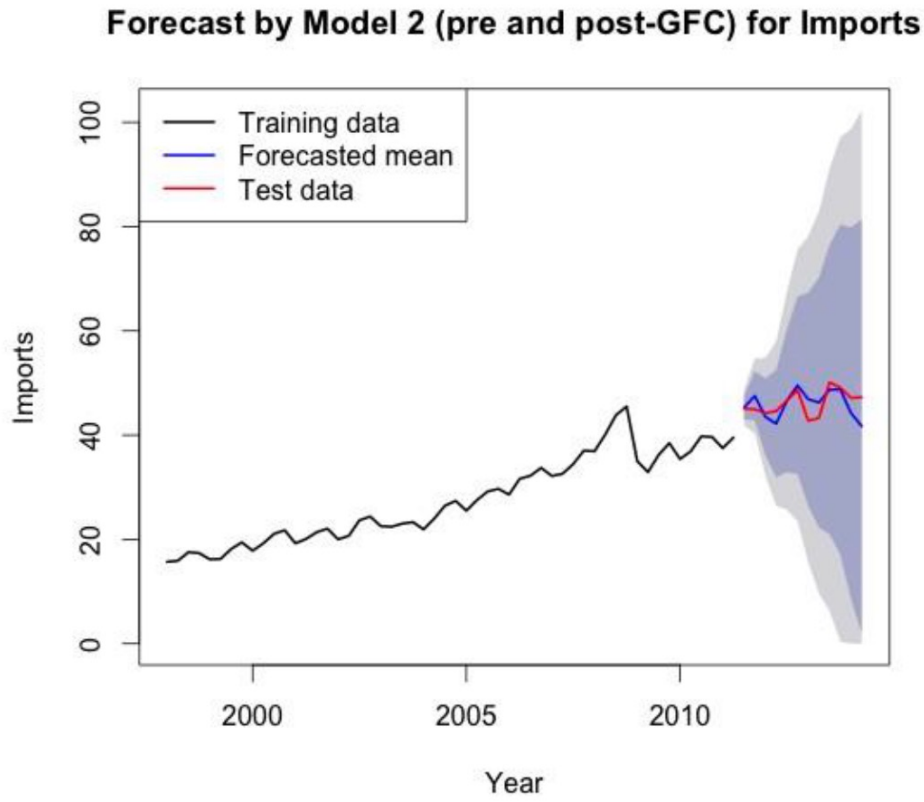
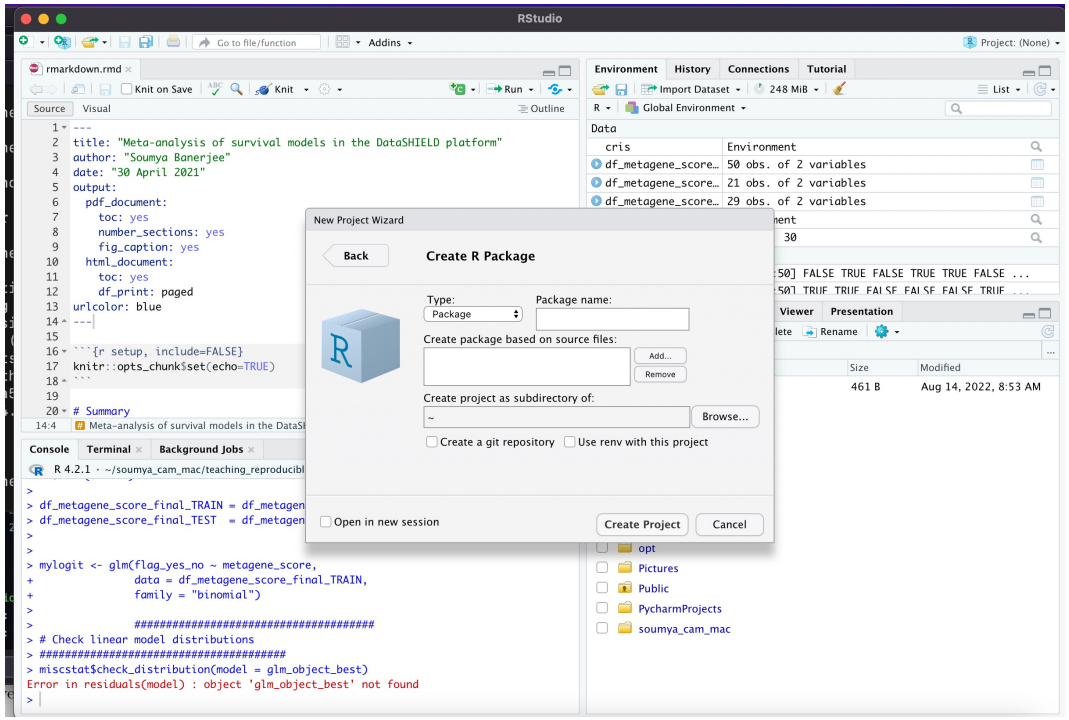


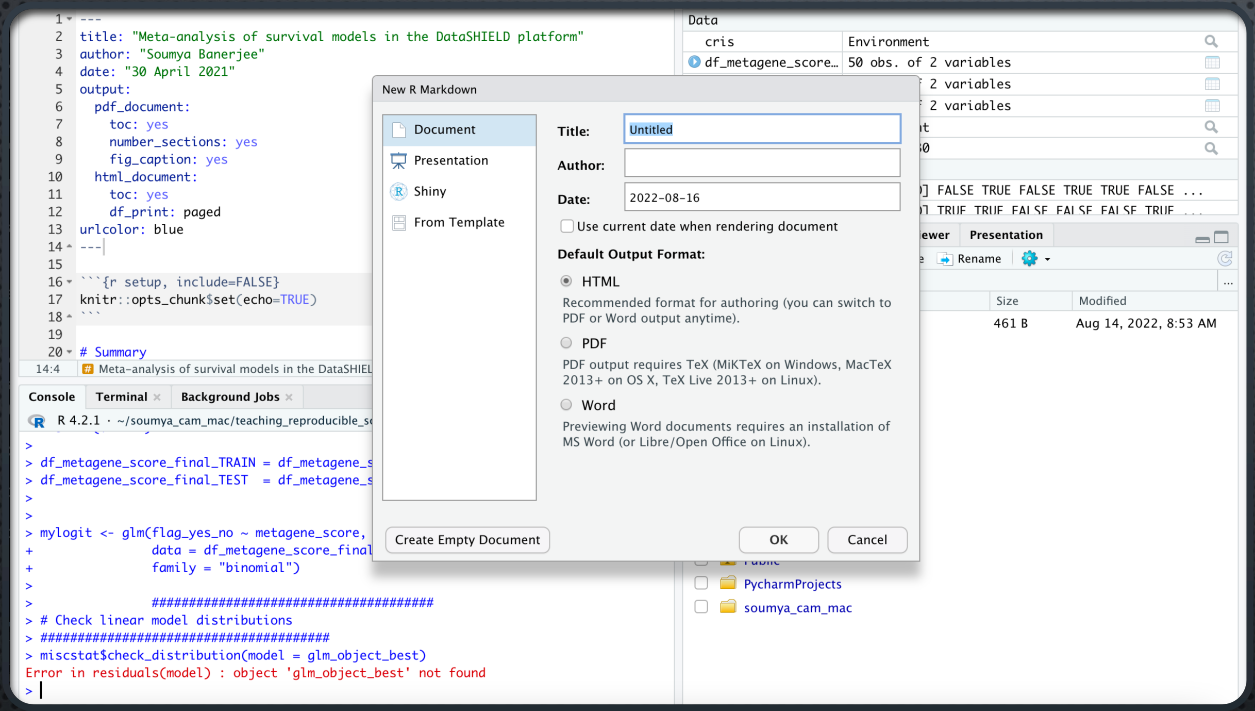

**Figure 7**. Forecast of port throughput by Model 2 (7D VAR model using GDPs of Australia and of top-5 importers as additional predictors of imports) trained on pre- and post-GFC data from Q1 1998 to Q2 2011. Out-of-sample (test set from Q3 2011 to Q2 2014) Root Mean Squared Error RMSE=2594.

# SHINY APPS FOR RAPID PROTOTYPING AND COMMUNICATION

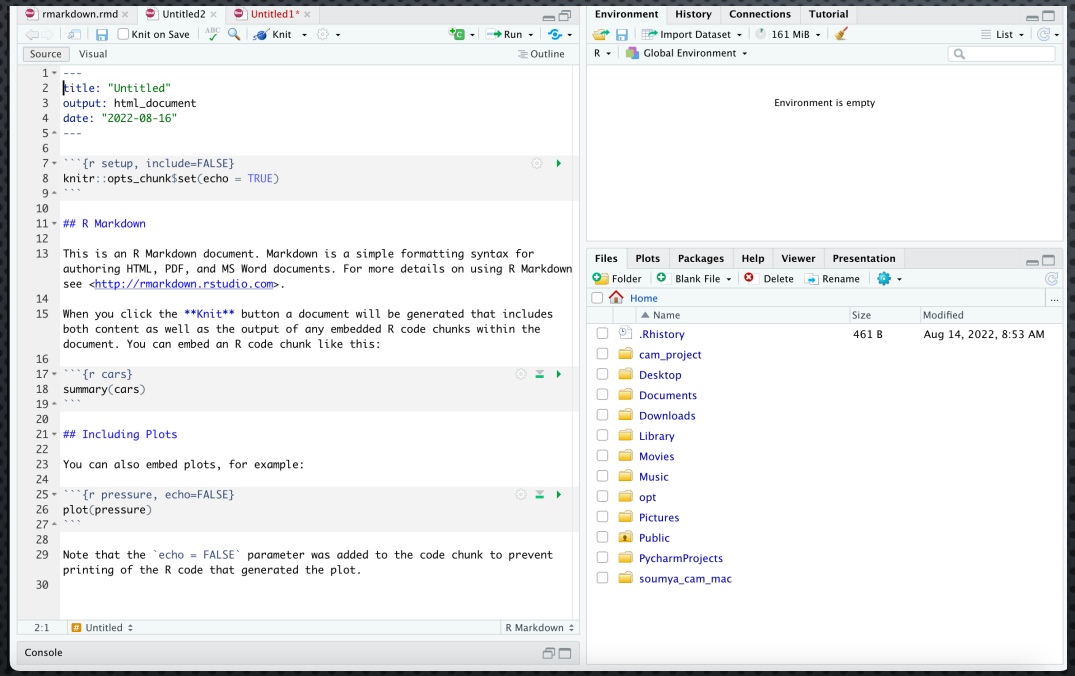
# SHINY APPS FOR RAPID PROTOTYPING AND COMMUNICATION




# CREATING PACKAGES




# CREATING R MARKDOWN


rmarkdown.rmd
Untitled2
Untitled1*
Knit on Save
Knit
Run
Source
Visual
Outline
---
title: "Untitled"
output: html_document
date: "2022-08-16"
---
```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```
## 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:
```{r cars}
summary(cars)
```
## Including Plots
You can also embed plots, for example:
```{r pressure, echo=FALSE}
plot(pressure)
```
Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.
2:1
Untitled
R Markdown
Console
Environment
History
Connections
Tutorial
Import Dataset
161 MiB
List
R
Global Environment
Environment is empty
Files
Plots
Packages
Help
Viewer
Presentation
Folder
Blank File
Delete
Rename
Home
Name
Size
Modified
.Rhistory
461 B
Aug 14, 2022, 8:53 AM
cam_project
Desktop
Documents
Downloads
Library
Movies
Music
opt
Pictures
Public
PycharmProjects
soumya_cam_mac

# DEMO

# 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 | Becomes |
| --- | --- |
| Make a code chunk with three back ticks followed by an r in braces. End the chunk with three back ticks:<br><br>```{r}<br>paste("Hello", "World!")<br>``` | Make a code chunk with three back ticks followed by an r in braces. End the chunk with three back ticks:<br><br>`paste("Hello", "World!")`<br><br>`## [1] "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!")`. | 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:<br><br>```{r eval=TRUE, echo=FALSE}<br>paste("Hello", "World!")<br>``` | Add chunk options within braces. For example, `echo=FALSE` will prevent source code from being displayed:<br><br>`## [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

- *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
- Now head over to the file named 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

# 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
- code 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/
- Code, resources and templates
- EXERCISES with synthetic data
- Installation instructions