## **Best Practices in Writing Reproducible Code in R**

Ms.C. Diego Angeles-Valdez University of Groningen | Instituto de Neurobiología, UNAM



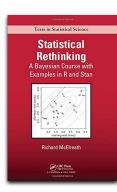


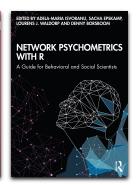
### **About me**













machine learning

Progress in Neuropsychopharmacology & Biological Psychiatry Identifying cognitive deficits in cocaine dependence using standard tests and

Said Jiménez<sup>1,5</sup>, Diego Angeles-Valdez<sup>1,5</sup>, Viviana Villicaña<sup>1,6</sup>, Ernesto Reyes-Zamorano<sup>1</sup>, Ruth Alcala-Lousno<sup>1</sup>, Jorge J. Goroalez-Overa<sup>1</sup>, Eduardo A. Garza-Villarreal<sup>1,6,6,6</sup>



Machine learning detects predictors of symptom severity and impulsivity after dialectical behavior therapy skills training group in borderline personality disorder

Said Jiménez<sup>\*\*</sup>, Diego Angeles-Valdez<sup>†</sup>, Andrés Rodríguez-Delgado<sup>\*</sup>, Ana Fresin <sup>†</sup>, Edgar Misanda<sup>\*</sup>, Ruth Alcalè-Lezano<sup>\*</sup>, Xóchill Duque-Alarcón<sup>\*</sup>, Iván Anango de Montis<sup>\*</sup>, Eduardo A. Gozza-Villarreal<sup>†, \*</sup>

1. In Normalization (170) is a relationar on distance (2011) fine from the Courte department (170) is a relationar on distance (2011) fine production (170) in a relationary of the Courte of the Cour

Found address regional/committed committee (EA Carus Villareed).

https://doi.org/10.2006/j.goptp.2019.100709
Received 28 January 2019; Received in revised from 29 July 2019; Accepted 22 July 2019
Armitalis cellus 25 July 2019
Armitalis cellus 25 July 2019
Armitalis cellus 25 July 2019

#### nature neuroscience

Received 8 May 2022

#### A consensus protocol for fun connectivity analysis in the r

on, resonance inagerig acquisioning. To develop this protocol with optimic testers, we initially aggregated 65 functions as across 46 centers. We developed a graz das acquired with diverse perior processing parameters associated we locomectivity across centres. We show thances biologically plausible function exercises acquisitions. The protocol and control of the protocol and the proto
re is openly shared with the neuroima ability and cooperation toward tackli neuroscience.

A list of authors and their affiliation

Understanding design design and the oliginate is more under the control of the co

First vs recurrent episode symptomatology in Major Depressive Disorder and its relation to brain function and structure; a network approach

Diego Angeles-Valdez \*\*\*\*, M.J. van Tol \*\*. Eduardo A. Garza-Villarreal \*\*

\* Institute of Neurobiology, Universidad Nacional Autónoma de México campus Juriquilla, Querétaro, México.

Cognitive Neuroscience Center, Department of Biomedical Sciences, University Medical Center Groningen, University of Groningen, The Netherlands.

6 Research School of Behavioural and Cognitive Neurosciences, University of Groningen, Groningen, The Netherlands.

#### \*Corresponding author:

#### Eduardo A. Garza-Villarreal, MD, PhD

Instituto de Neurobiología, Laboratorio D-12, Universidad Nacional Autónoma de México (UNAM) campus Juriquilla, Boulevard Juriguilla 3001, Santiago de Querétaro, Querétaro, México

Email: egarza@comunidad.unam.mx

#### Phone: (442) 238-1038 Diego Angeles-Valdez, M. Sc.

PhD candidate. Instituto de Neurobiología, Laboratorio D-12, Universidad Nacional Autónoma de México (UNAM) campus Juriquilla, University Medical Center Groningen, Cognitive Neuroscience Center, Hanzeplein 1, 9713 GZ Groningen

y & Biological Parchisery 129 (2025) 111397

of construct (E.A. Garra-Villamod).

opharmacology & Biological chiatry

groups in the Intermittent-Access at class linear mixed model

pulsate respond to therapy, and the finatures or variables that predict the clinical response are still unknown ("climitizen et al., 2005). Me repose that the folialisation of their perfectors can largown efficient outcome proposels and postatisfy increase the number of patients who unknown proposels and postatisfy increase the number of patients who have been predicted as the proposel parties with a patient patient patient patients of the patient patients of the patients o

, Jalil Rasgado-Toledo ,

public bash common (bearing of a 2013). Solidation dissulted (1003) in district to consumer other over or channel from two derivation of the other of the Channel of the Internal Marinel of Marinel Channel of the points of dealed see that condense described and the Internal Marinel of the Internal Marinel Marinel of the Internal Marinel Mari

https://doi.org/10.2105/j.pupin.2005.11.2007

Retrieved 19 February 2005. Retrieved is revised from 24 April 2020; Accepted 5 May 2020

Available under 16 Very 2020

Available under 16 Very 2020

Available to be also. To Additional Systems of the April 2020; Accepted 5 May 2020

Available to Be April 2020

Av

### **Outline**



### R Packages

Overview of essential R packages used in data analysis

### Glimpse of the Data

Initial exploration and summary of data

### **Statistical Analysis**

Applying statistical methods to interpret data

#### **Visualization**

Creating visual representations to illustrate findings

### **Course Material**





*Tidyverse* 

*Tidy* for "bien rangé" and *verse* for "univers"

A collection of R will developed by H. Wickham and others at Rstudio

Initial release: Sep, 2016



**Hadley Wickham** 



Tidyverse is most importantly a philosophy for data analysis.

- More efficient code
- Easier to remember syntax
- Easier to read syntax



### Remember install and load the package

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



```
library(tidyverse)
                                                                tidyverse 1.3.0 —
## — Attaching packages
## / ggplot2 3.3.3
                                0.3.4
                      ✓ purrr
## ✓ tibble 3.1.0

✓ dplyr 1.0.4

## / tidyr 1.1.2
                      ✓ forcats 0.5.1
## / readr 1.4.0
## — Conflicts
                                                         - tidyverse conflicts() —
## * dplyr::collapse()
                        masks IRanges::collapse()
## * dplyr::combine()
                        masks Biobase::combine(), BiocGenerics::combine()
## * dplyr::desc()
                        masks IRanges::desc()
## * tidyr::expand()
                        masks S4Vectors::expand()
## * tidyr::extract()
                        masks magrittr::extract()
## * dplyr::filter()
                        masks stats::filter()
                        masks S4Vectors::first()
## * dplyr::first()
## * dplyr::lag()
                        masks stats::lag()
## * ggplot2::Position() masks BiocGenerics::Position(), base::Position()
## * purrr::reduce()
                        masks GenomicRanges::reduce(), IRanges::reduce()
## * dplyr::rename()
                        masks S4Vectors::rename()
```



Tidyverse is a collection of R



ggplot2 - visualising stuff

**dplyr**, **tidyr** - data manipulation

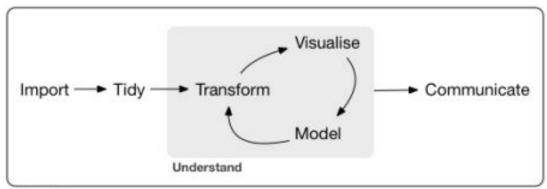
purrr - advanced programming

readr - import data

**tibble** - improved data.frame format

forcats - working w/ factors

**stringr** - working w/ chain of characters



Program

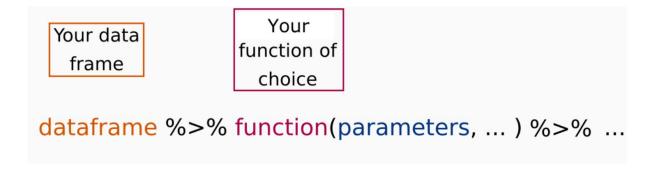
## Introduction to dplyr

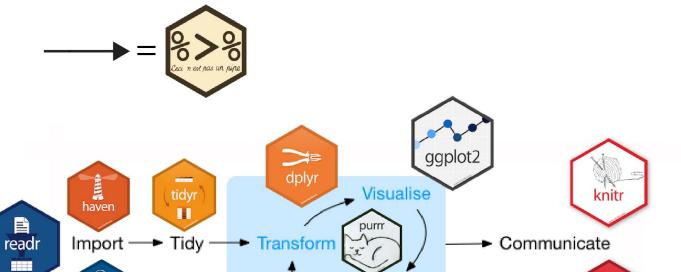


All of the dplyr functions take a data frame (or tibble) as the first argument.

dplyr provides the %>%

You can use the pipe to rewrite multiple operations that you can read left-to-right, top-to-bottom.



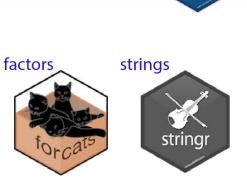


Explore

readxl

data.frames

7133LE



Model

broom

rmarkdown

# Graphics in R



The ggplot2 package was created by Hadley Wickham to provide an intuitive plotting system.

In order to produce a ggplot2 graph we need a minimum of:

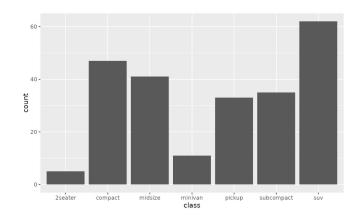
- Data to be used in graph
- Mappings of data to the graph (aesthetic mapping)
- What type of graph we want to use (The geom to use).

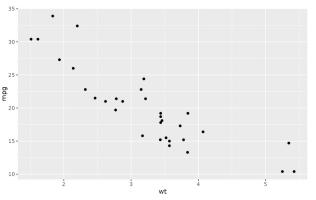
## Our first ggplot2 graph

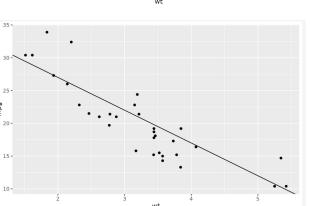


### For example:

- geom\_point() for scatter plots,
- geom\_line() for line charts,
- geom\_bar() for bar plots,
- geom\_histogram() for histograms,
- geom\_smooth() to add trend lines.



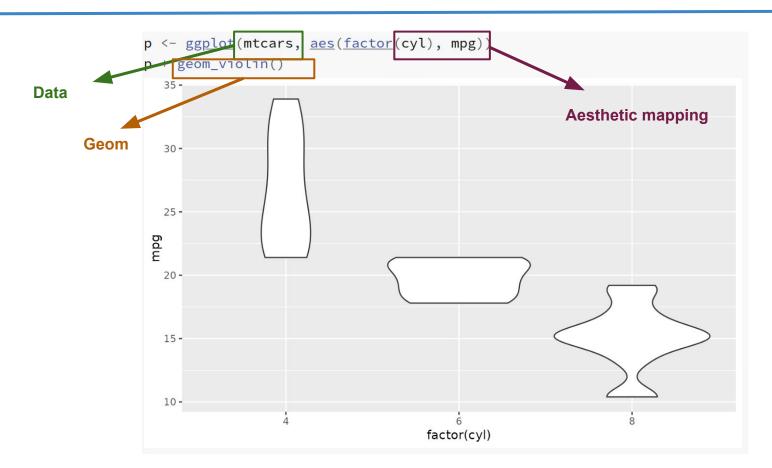






# Our first ggplot2 graph





### **Table**



Descriptive statistics are an essential part of data analysis as they give you an initial overview of your data.

To create descriptive tables in R, you can use:

- moonBook
- table1
- tableone
- tables

# **Examples**



	Stratifie	by trt				
	1		2		p	tes
n	158		154			
time (mean (SD))	2,015.62	(1,094.12)	1,996.86	(1,155.93)	0.883	
status (%)					0.894	
0		(52.5)		(55.2)		
1	10	(6.3)		( 5.8)		
2		(41.1)	60	(39.0)		
trt = 2 (%)	0	(0.0)	154	(100.0)	<0.001	
age (mean (SD))	51.42	(11.01)	48.58	(9.96)	0.018	
sex = f (%)	137	(86.7)	139	(90.3)	0.421	
ascites = 1 (%)	14	(8.9)	10	( 6.5)	0.567	
hepato = 1 (%)	73	(46.2)	87	(56.5)	0.088	
spiders = 1 (%)	45	(28.5)	45	(29.2)	0.985	
edema (%)					0.877	
0	132	(83.5)	131	(85.1)		
0.5	16	(10.1)	13	(8.4)		
1	10	(6.3)	10	(6.5)		
bili (median [IQR])	1.40	[0.80, 3.20]	1.30	[0.72, 3.60]	0.842	
nonnorm						
chol (median [IQR])	315.50	[247.75, 417.00]	303.50	[254.25, 377.00]	0.544	
nonnorm						
albumin (mean (SD))	3.52	(0.44)	3.52	(0.40)	0.874	
copper (median [IQR])	73.00	[40.00, 121.00]	73.00	[43.00, 139.00]	0.717	
nonnorm						
alk.phos (median [IQR])	1,214.50	[840.75, 2,028.00]	1,283.00	[922.50, 1,949.75]	0.812	
nonnorm						
ast (median [IQR])	111.60	[76.73, 151.51]	117.40	[83.78, 151.90]	0.459	
nonnorm		[		[		
trig (median [IQR])	106.00	[84.50, 146.00]	113.00	[84.50, 155.00]	0.370	
nonnorm		[,]		[,		
platelet (mean (SD))	258 75	(100 32)	265 20	(90.73)	0.555	
protime (median [IQR])					0.588	
nonnorm	10.00	[10.00, 11.00]	10.00	[20.00, 11.70]	0.000	
stage (%)					0.201	
1	10	(7.6)	А	( 2.6)	0.201	
2				( 20.8)		
		(22.2)				
3		(35.4)		(41.6)		
4	55	(34.8)	54	( 35.1)		

	NSTEMI (N=153)		Unstable Angina (N=400)	р	
age	64.3 ± 12.3	62.1 ± 12.1	63.8 ± 11.0	0.073	
sex				0.013	
- Female	50 (32.7%)	84 (27.6%)	153 (38.2%)		
- Male	103 (67.3%)	220 (72.4%)	247 (61.8%)		
cardiogenicShoc	k			0.000	
- No	149 (97.4%)	256 (84.2%)	400 (100.0%)		
- Yes	4 ( 2.6%)	48 (15.8%)	0 (0.0%)		
entry				0.00	
- Femoral	58 (37.9%)	133 (43.8%)	121 (30.2%)		
- Radial	95 (62.1%)	171 (56.2%)	279 (69.8%)		
EF	55.0 ± 9.3	52.4 ± 9.5	59.2 ± 8.7	0.00	
height	163.3 ± 8.2	165.1 ± 8.2	161.7 ± 9.7	0.000	
weight	64.3 ± 10.2	65.7 ± 11.6	64.5 ± 11.6	0.36	
BMI	24.1 ± 3.2	24.0 ± 3.3	24.6 ± 3.4	0.06	
obesity				0.18	
- No	106 (69.3%)	209 (68.8%)	252 (63.0%)		
- Yes	47 (30.7%)	95 (31.2%)	148 (37.0%)		
TC	193.7 ± 53.6	183.2 ± 43.4	183.5 ± 48.3	0.05	
LDLC	126.1 ± 44.7	116.7 ± 39.5	112.9 ± 40.4	0.004	
HDLC	38.9 ± 11.9	38.5 ± 11.0	37.8 ± 10.9	0.50	
TG	130.1 ± 88.5	106.5 ± 72.0	137.4 ± 101.6	0.00	
DM				0.20	
- No	96 (62.7%)	208 (68.4%)	249 (62.2%)		
- Yes	57 (37.3%)	96 (31.6%)	151 (37.8%)		
HBP	7.CA - 7.24 - 7.24 - 7.24 - 7.24	A CONTRACTOR OF THE CONTRACTOR	are an area of the same	0.002	
- No	62 (40.5%)	150 (49.3%)	144 (36.0%)		
- Yes		154 (50.7%)			
smoking			**************************************	0.00	
- Ex-smoker	42 (27.5%)	66 (21.7%)	96 (24.0%)		
- Never		97 (31.9%)			
- Smoker		141 (46.4%)			

#### Basic stats

	Dusic stats				
			·	Death	
	Total (N=205)	Alive (N=134)	Melanoma (N=57)	Non-melanoma (N=14)	
Sex					
Male	79 (39 %)	43 (32 %)	29 (51 %)	7 (50 %)	
Female	126 (61 %)	91 (68 %)	28 (49 %)	7 (50 %)	
Age (years)					
Mean (SD)	52 (± 17)	50 (± 16)	55 (± 18)	65 (± 11)	
Ulceration					
Absent	115 (56 %)	92 (69 %)	16 (28 %)	7 (50 %)	
Present	90 (44 %)	42 (31 %)	41 (72 %)	7 (50 %)	
Thicknessa (mm)					
Mean (SD)	2.9 (± 3.0)	2.2 (± 2.3)	4.3 (± 3.6)	3.7 (± 3.6)	

Also known as Breslow thickness

		Sepal. Length		Sepal.Width		
Species	$\mathbf{n}$	mean	sd	mean	$_{ m sd}$	
setos a	50	5.01	0.35	3.43	0.38	
versicolor	50	5.94	0.52	2.77	0.31	
virginica	50	6.59	0.64	2.97	0.32	
Overall, w	e see i	the follo	wing:			
All	150	5.84	0.83	3.06	0.44	

# ggplot2



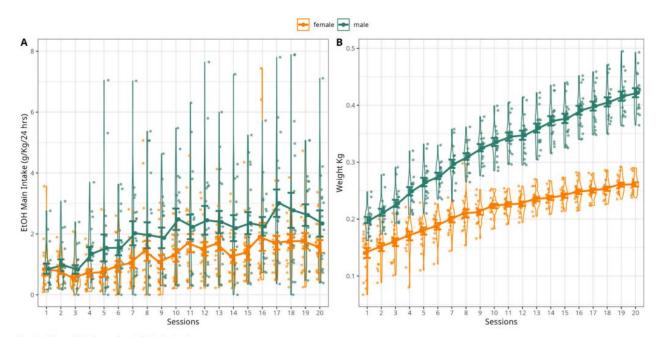




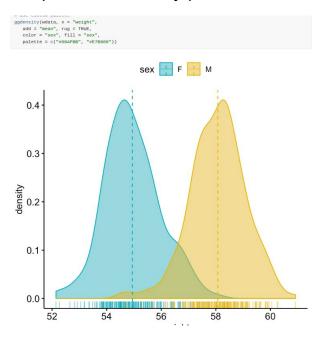
Fig. 1. Ethanol intake and weight trajectories.

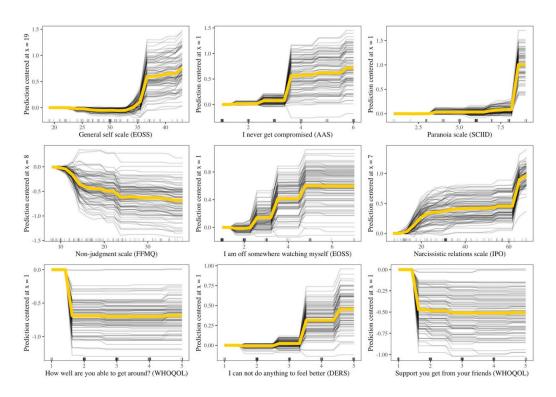
Overview of (A) ethanol main intake, (B) and the weight of all rats over IA2BC model by sex. The values are expressed as mean EtOH Main intake (g/kg/24 h), and weight (Kg)  $\pm$  SEM at each drinking session.

## ggpubr



The 'ggpubr' package provides some easy-to-use functions for creating and customizing 'ggplot2'-based publication ready plots.





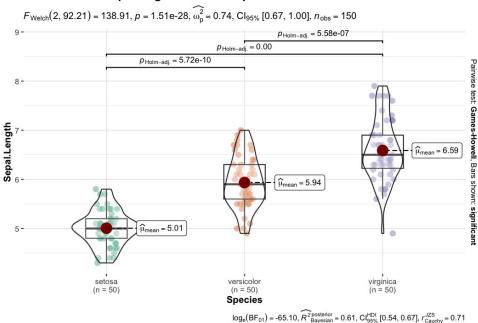
# Other R packages for plotting





Based Plots with Statistical Details

#### Distribution of sepal length across Iris species



### Other R packages for plotting



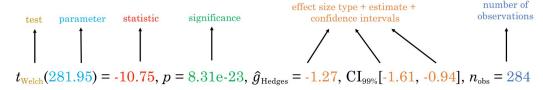


Based Plots with Statistical Details

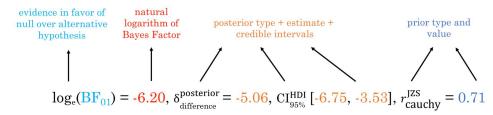


Results from Welch's t-test with {statsExpressions}

#### Template for Frequentist analysis



#### Template for Bayesian analysis



### **Extra information**



### Linear Mixed models Network analysis

- 1mer
- broom

- SEM or path model
  - lavaan

- bootnet
- psychonetrics
- qgraph

### Machine Learning

- caret
- randomForest
- o xgboost
- o e1071
- tidymodels

# Thank you



### Contact me

M. Sc. Diego Angeles-Valdez, PhD candidate,

University Medical Center Groningen, Cognitive Neuroscience Center, Biomedical Sciences

> Universidad Nacional Autónoma de México (UNAM) Institute of Neurobiology (INB)

Personal Website: <a href="https://diegoangls.github.io/">https://diegoangls.github.io/</a>

Twitter: <a>@diegoangls</a>

Bluesky: @dangeles.bsky.social

Phone: +31 6 39 88 71 29

Email: d.angeles.valdez@rug.nl | d.angeles.valdez@umcg.nl

Work address: P.O. Box 196, 9700 AD GRONINGEN, The Netherlands