

@Seminar{

title = {Levelling up R for Statistical Research and
Teaching},

author = {Aleeza Gerstein},

year = {2019},

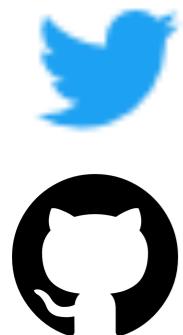
note = {UManitoba Statistics Department},

url = {github.com/acgerstein/seminars/}

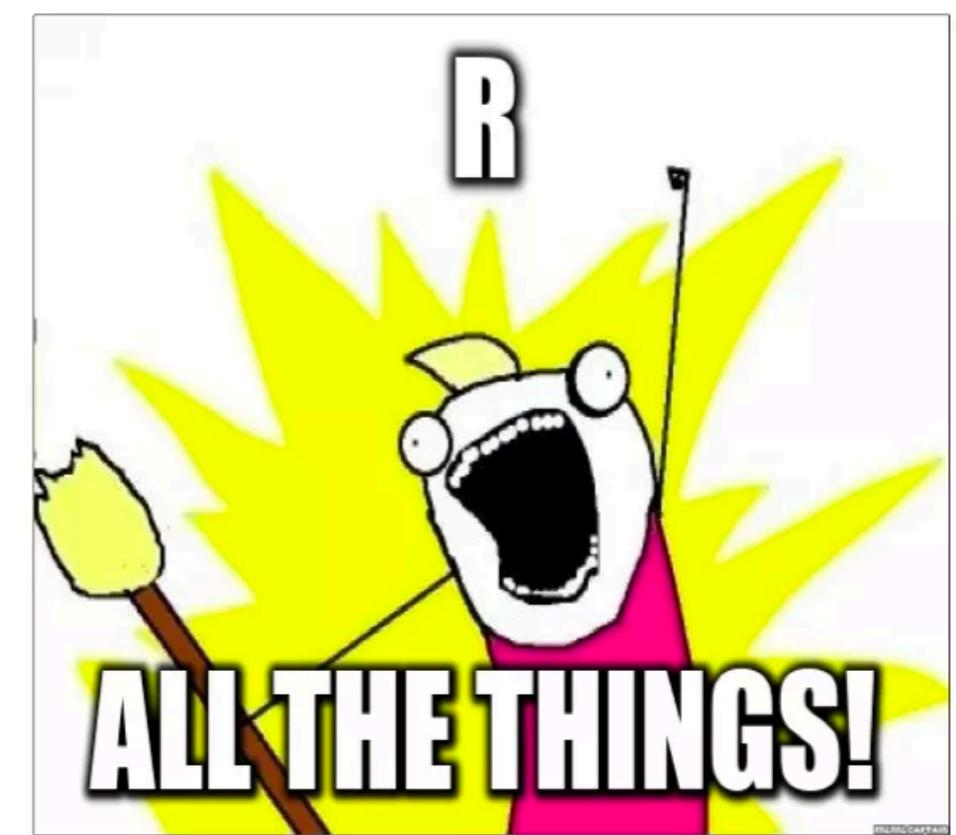
2019RAIITheThings.pdf},

}

@acgerstein

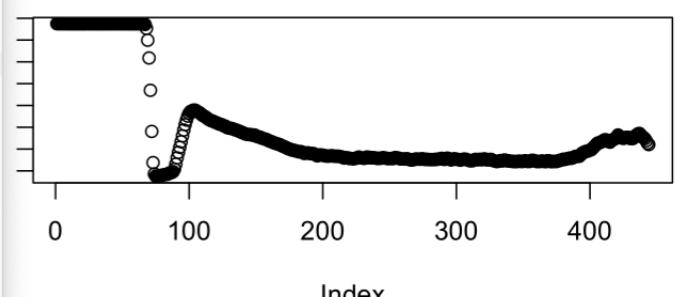
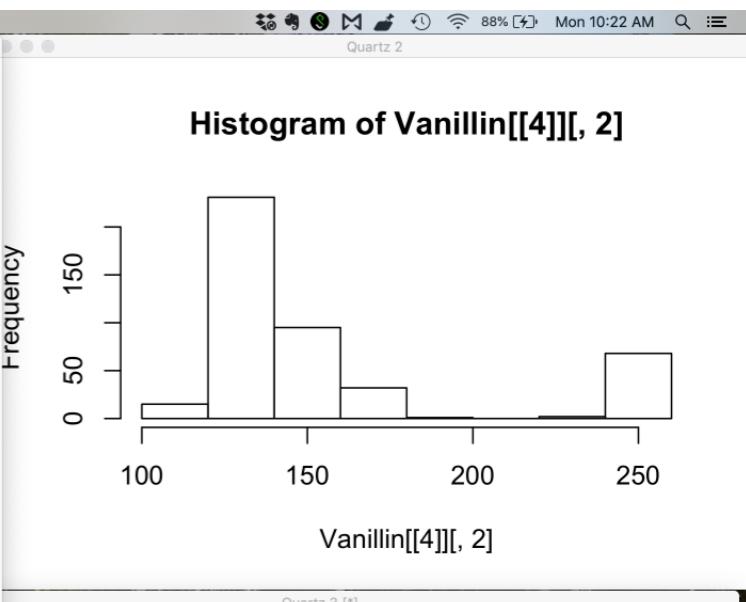


faculty of SCIENCE
discover the unknown + invent the future



install.packages("meme")

This is how I R (since 2004)



```
diskImageR --bash -- 80x24
Documents Library Pictures
[scist-mh364-584:~ agerstein$ cd Documents/Postdoc/Research/diskImageR
-bash: cd: command not found
[scist-mh364-584:~ agerstein$ cd Documents/Postdoc/Research/diskImageR/
-bash: cd: command not found
[scist-mh364-584:~ agerstein$ cd Documents/Postdoc/Research/diskImageR/
[scist-mh364-584:~ diskImageR agerstein$ ls
DESCRIPTION R diskImageR man
NAMESPACE README.md inst vignettes
[scist-mh364-584:~ diskImageR agerstein$ git add R/maxlik.R
[scist-mh364-584:~ diskImageR agerstein$ git commit -m "add plotCompon to plotIndiv"
[master b62a64f] add plotCompon to plotIndiv
 1 file changed, 48 insertions(+), 43 deletions(-)
[scist-mh364-584:~ diskImageR agerstein$ git push
Counting objects: 4, done.
Delta compression using up to 8 threads.
Compressing objects: 100% (4/4), done.
Writing objects: 100% (4/4), 697 bytes | 697.00 KiB/s, done.
Total 4 (delta 3), reused 0 (delta 0)
remote: Resolving deltas: 100% (3/3), completed with 3 local objects.
To https://github.com/agerstein/diskImageR.git
 27339fc..b62a64f master -> master
[scist-mh364-584:~ diskImageR agerstein$ ]
```

This is how I R (since 2004)

Base console

Base plot

Git terminal

The image shows four windows illustrating the R environment from 2004:

- Base console:** An R console window showing command-line interactions and output. A large blue arrow points from the console up towards the plot window.
- Base plot:** A histogram titled "Histogram of Vanillin[[4]][, 2]" showing the frequency distribution of Vanillin values. The x-axis is labeled "Vanillin[[4]][, 2]" and ranges from 100 to 250. The y-axis is labeled "Frequency" and ranges from 0 to 150. A large blue arrow points from the console down towards the plot window.
- Git terminal:** A terminal window showing a git session for the "diskImageR" repository. It includes commands like "git commit", "git push", and "git pull". A large blue arrow points from the plot window down towards the terminal window.
- R Console:** The main R console window showing a complex R script named "maxLik.R". The script contains numerous lines of R code, including functions like "singlePlot" and "inhibGrowPts.R". The R interface includes a menu bar (Finder, File, Edit, View, Go, Window, Help) and a toolbar with various icons.

This is how my students R

RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help

~/Documents/Umanitoba/Courses/data-science-in-tidyverse - master - RStudio

RMfigures.R x

```
11 names(df)[2] <- "species"
12
13 labels <-
14 expression(
15   italic("C. albicans"), italic("C. glabrata"))
16
17
18 pal <- wes_palette("FantasticFox1")
19 pal[1]
20
21 ggplot(df, aes(study, Percent, fill = species)) +
22   geom_col(position = "dodge") +
23   scale_fill_manual(values = c(pal[3], pal[2], pal[5]), labels=c(expression(italic("C. albicans")), italic("C. glab:
24   theme_bw(base_size = 20) +
25   labs (y = "Percent (%)") +
26   theme(legend.text.align = 0) +
27   theme(legend.position = "top", legend.title = element_text(size = 14), legend.text = element_text(size = 12))
28
```

19:7 (Top Level) R Script

Console Terminal x R Markdown x

```
2   B  50   35   15
3   C  69   23    8
4   D  76    9   15
5   E  81   18    1
6   F  95    5    0
>
> df <- melt(df, value.name = "Percent")
Using study as id variables
> names(df)[2] <- "species"
>
> labels <-
+ expression(
+   italic("C. albicans"), italic("C. glabrata"))
>
> pal <- wes_palette("FantasticFox1")
> pal[1]
[1] "#DD8D29"
>
> ggplot(df, aes(study, Percent, fill = species)) +
+   geom_col(position = "dodge") +
+   scale_fill_manual(values = c(pal[3], pal[2], pal[5]), labels=c(expression(italic("C. albicans")), italic("C. glab:
+   theme_bw(base_size = 20) +
+   labs (y = "Percent (%)") +
+   theme(legend.text.align = 0) +
+   theme(legend.position = "top", legend.title = element_text(size = 14), legend.text = element_text(size = 12))
>
```

Notes

Environment History Connections Git

Diff Commit Pull Push History More New Branch master

Staged Status Path 01-Visualize.Rmd

Files Plots Packages Help Viewer

Zoom Export

species C. albicans C. glabrata other

Percent (%)

study

study	C. albicans	C. glabrata	other
A	~68%	~20%	~12%
B	~50%	~35%	~15%
C	~70%	~25%	~10%
D	~75%	~10%	~15%
E	~78%	~15%	~2%
F	~80%	~5%	~0%

This is how my students R

RStudio

Integrated git

ggplot

The screenshot shows the RStudio interface with several key components highlighted:

- RStudio:** The main menu bar at the top.
- Integrated git:** The Git tab in the top right corner, showing a local repository for "data-science-in-tidyverse".
- Console:** The bottom-left pane showing R code and its output. A large blue arrow points from the word "RStudio" to the console area.
- Plots:** The bottom-right pane displaying a grouped bar chart titled "Percent (%)". The x-axis is labeled "study" and has categories A, B, C, D, E, F. The y-axis ranges from 0 to 75. The legend indicates three species: C. albicans (blue), C. glabrata (yellow), and other (red). The chart shows varying proportions across studies, with C. albicans being dominant in most cases.
- Code Editor:** The top-left pane showing an R script named "RMfigures.R". A large blue arrow points from the word "Integrated git" to the code editor area.

RStudio (the IDE & company) is the future

R: Engine



RStudio: Dashboard



<http://moderndive.com/>



This is how we could teach R

The screenshot shows the RStudio Cloud interface running in a Chrome browser window. The main area displays an R script titled "LearningToolsWeek01.R" and its corresponding output in the console. The script performs statistical calculations, including the use of the dpois function to calculate probabilities for a Poisson distribution with lambda = 4.21 across 20 categories (0:20). It then calculates expected probability, length, and combined values. The environment pane shows the current workspace, and the file browser pane shows the project structure under "Cloud > project > ABDLabs".

R Script Content:

```
dpois(x = 3, lambda = 4.21)
0:20
expected_probability = dpois(x = 0:20, lambda = 4.21)
expected_probability
length(number_of_extinctions)
76 * expected_probability
expected_combined <- c(5.878568, 9.999264, 14.032300, 14.768996, 12.435494, 8.725572, 5.247808, 4.911998)
observed_combined <- c(13, 15, 16, 7, 10, 4, 2, 9)
```

Console Output:

```
> 0:20
[1] 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
> expected_probability = dpois(x = 0:20, lambda = 4.21)
> expected_probability
[1] 1.484637e-02 6.250321e-02 1.315693e-01 1.846355e-01 1.943289e-01 1.636249e-01
[7] 1.148102e-01 6.905011e-02 3.633762e-02 1.699793e-02 7.156129e-03 2.738846e-03
[13] 9.608784e-04 3.111768e-04 9.357530e-05 2.626347e-05 6.910575e-06 1.711384e-06
[19] 4.002736e-07 8.869220e-08 1.866971e-08
> length(number_of_extinctions)
[1] 76
>
> 76 * expected_probability
[1] 1.128324e+00 4.750244e+00 9.999264e+00 1.403230e+01 1.476900e+01 1.243549e+01
[7] 8.725572e+00 5.247808e+00 2.761659e+00 1.291843e+00 5.438658e-01 2.081523e-01
[13] 7.302676e-02 2.364943e-02 7.111723e-03 1.996023e-03 5.252037e-04 1.300651e-04
[19] 3.042079e-05 6.740607e-06 1.418898e-06
>
> expected_combined <- c(5.878568, 9.999264, 14.032300, 14.768996, 12.435494, 8.725572, 5.247808, 4.911998)
> observed_combined <- c(13, 15, 16, 7, 10, 4, 2, 9)
>
```

This is how we could teach R

A screenshot of a Mac OS X desktop showing the RStudio Cloud interface running in a Chrome browser window. A large blue arrow points from the top left towards the browser's address bar, which displays the URL <https://rstudio.cloud/project/232512>. The RStudio Cloud interface includes a sidebar with 'Spaces' (Your Workspace selected), 'Learn' (Guide, What's New, Primers, DataCamp Courses, Cheat Sheets, Feedback and Questions), and 'Info' (Terms and Conditions, System Status). The main workspace shows an R script editor with code for calculating expected probabilities and combined values, and a terminal window showing the same calculations. A file browser on the right lists files in a project named 'ABDLabs'. A prominent black-bordered box contains the text 'RStudio Cloud through the web!'.

RStudio Cloud through the web!

```
52 dpois(x = 3, lambda = 4.21)
53
54 0:20
55
56 expected_probability = dpois(x = 0:20, lambda = 4.21)
57 expected_probability
58
59 length(number_of_extinctions)
60
61 76 * expected_probability
62
63 expected_combined <- c(5.878568, 9.999264, 14.032300, 14.768996, 12.435494, 8.725572, 5.247808, 4.911998)
64 observed_combined <- c(13, 15, 16, 7, 10, 4, 2, 9)
65
66 chi_squared_statistic
67
68 # Analysis of Biological Data Labs -- Learning the tools -- Week 6
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
```

A very brief history

S



1976

Rick Becker and John Chambers

1996

Ross Ihaka and Robert Gentleman
[R Foundation in 2003]



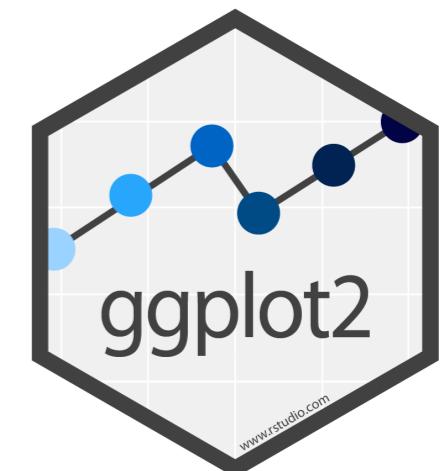
Bell Laboratories



Carlos Grajales: <http://www.statisticsviews.com/details/feature/8585391/Created-by-statisticians-for-statisticians-How-R-took-the-world-of-statistics-by.html>

A very brief history

S



1976

Rick Becker and John Chambers

1996

Ross Ihaka and Robert Gentleman
[R Foundation in 2003]

2005

Hadley Wickham

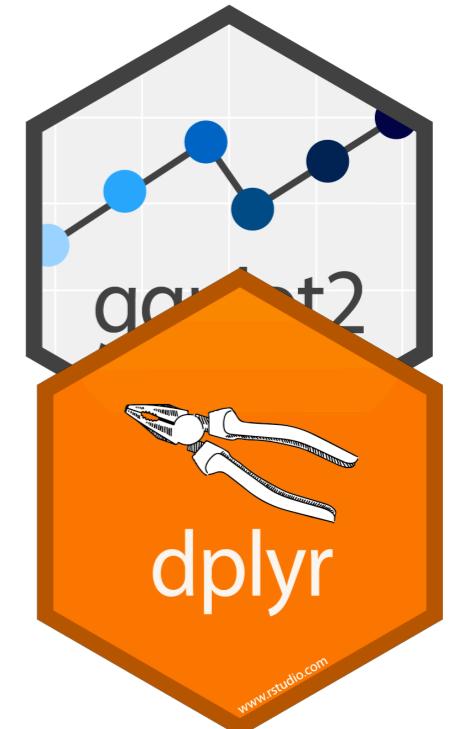


Bell Laboratories



A very brief history

S



1976

Rick Becker and John Chambers

1996

Ross Ihaka and Robert Gentleman
[R Foundation in 2003]

2013

Hadley Wickham



Bell Laboratories



A very brief history

S



1976

Rick Becker and John Chambers

1996

Ross Ihaka and Robert Gentleman
[R Foundation in 2003]

2015
RStudio



Bell Laboratories

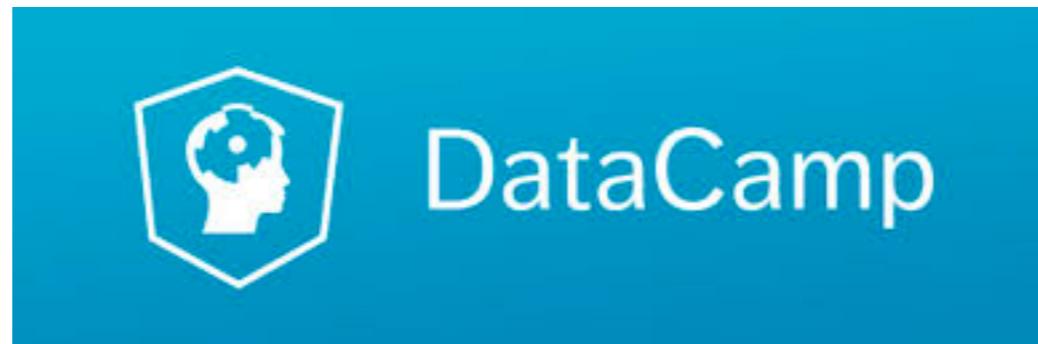
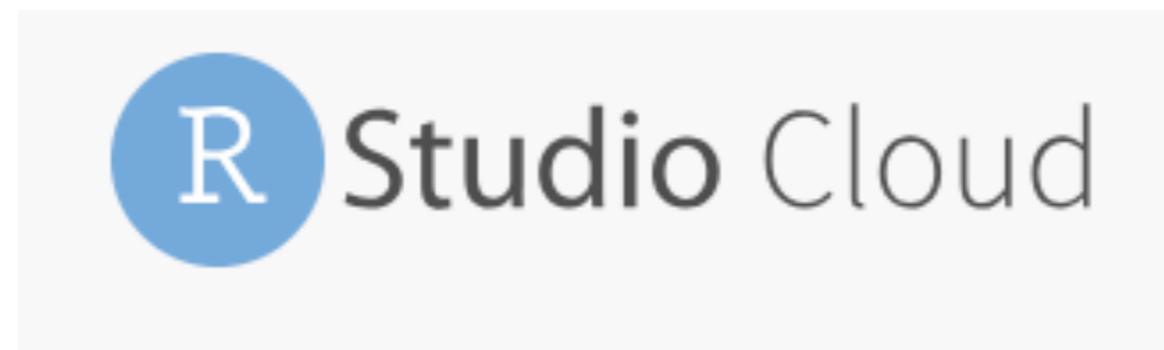


Carlos Grajales: <http://www.statisticsviews.com/details/feature/8585391/Created-by-statisticians-for-statisticians-How-R-took-the-world-of-statistics-by.html>

Talk outline



vs.





- software environment for statistical computing and graphics.
- flexible and programmable
- total control to user

Roger Peng: <https://simplystatistics.org/2018/07/12/use-r-keynote-2018/>



packages for everything

[car](#) – car's Anova function is popular for making type II and type III Anova tables.

[mgcv](#) – Generalized Additive Models

[lme4/nlme](#) – Linear and Non-linear mixed effects models

[randomForest](#) – Random forest methods from machine learning

[multcomp](#) – Tools for multiple comparison testing

[vcd](#) – Visualization tools and tests for categorical data

[glmnet](#) – Lasso and elastic-net regression methods with cross validation

[survival](#) – Tools for survival analysis

[caret](#) – Tools for training regression and classification models

Hosted on CRAN, bioconductor, GitHub



vs.



- software environment for statistical computing and graphics.
- flexible and programmable
- total control to user
- opinionated collection of R packages designed for data science ("metapackage")
- designed for users with no previous programming experience, encountering data science and statistics for the first time
- guides users through workflows that improve reproducibility and communication

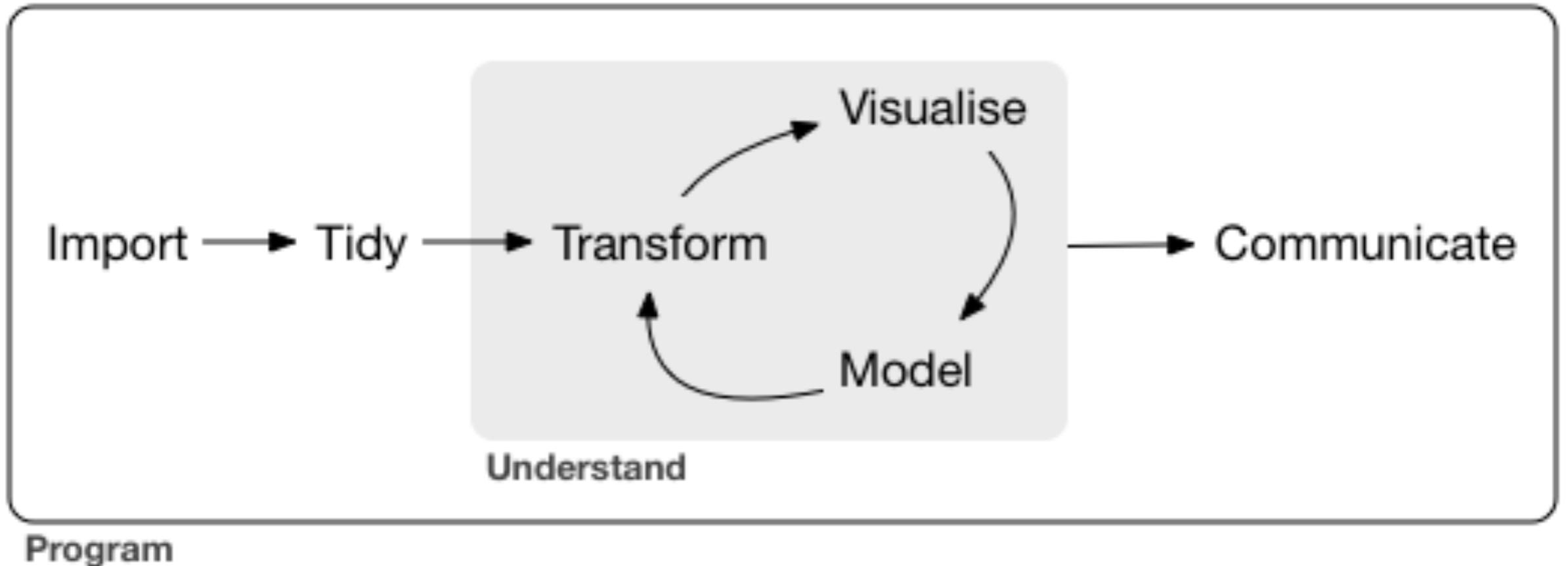


```
aggregate(airquality[, "Ozone"],  
         list(Month = airquality[, "Month"] ),  
         mean, na.rm = TRUE)
```

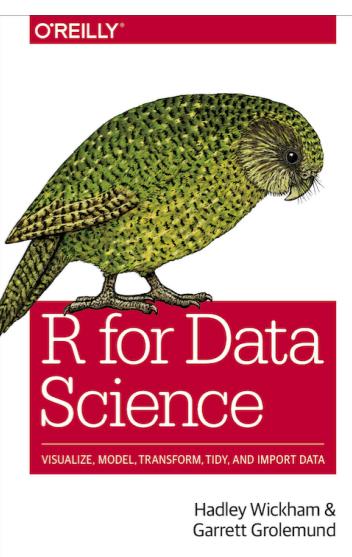


```
group_by(airquality, Month) %>%  
  summarize(o3 = mean(Ozone, na.rm = TRUE))
```

Tidyverse workflow



Hadley Wickham: <https://r4ds.had.co.nz/>
Tidy Data. Journal of Statistical Software, Vol. 59, No. 10.



Consistency



- **Variables, functions, operators follow regular patterns and syntax**
 - First argument of a function is always a tidy data frame

baker	cinnamon_1	cardamom_2	nutmeg_3
Emma	1	0	1
Harry	1	1	1
Ruby	1	0	1
Zainab	0	NA	0

baker	spice	correct
Emma	cinnamon_1	1
Harry	cinnamon_1	1
Ruby	cinnamon_1	1
Zainab	cinnamon_1	0
Emma	cardamom_2	0
Harry	cardamom_2	1
Ruby	cardamom_2	0
Zainab	cardamom_2	NA
Emma	nutmeg_3	1
Harry	nutmeg_3	1
Ruby	nutmeg_3	1
Zainab	nutmeg_3	0



Alison Hill
@rstudio

Readability



- **Pipe allows operations coded the way you read**

- no nested function calls
- no intermediate steps
- minimize creation of local variables and overwriting original data

<https://www.r-bloggers.com/readable-code-with-pipes/>

Readability



- Pipe allows operations coded the way you read

```
> head(ecom2)
  referrer n_pages duration purchase
1 google      1       693   false
2 yahoo       1       459   false
3 direct      1       996   false
4 bing        18      468    true
5 yahoo       1       955   false
6 yahoo       5       135   false
.
```

```
| > y <- sqrt(ecom2$n_pages)
```

```
| > y <-
| + ecom2$n_pages %>%
| + sqrt()
```

Readability



- Pipe allows operations coded the way you read

```
> head(ecom)
  referrer n_pages duration purchase
1 google      1       693   false
2 yahoo       1       459   false
3 direct      1       996   false
4 bing        18      468    true
5 yahoo       1       955   false
6 yahoo       5       135   false
```

```
> ecom1 <- subset(ecom, purchase)
> cor(ecom1$n_pages, ecom1$duration)
[1] 0.4290905
```

```
> ecom %>%
+   subset(purchase) %>%
+   cor(n_pages, duration)
[1] 0.4290905
```

Readability

- Function names are verbs

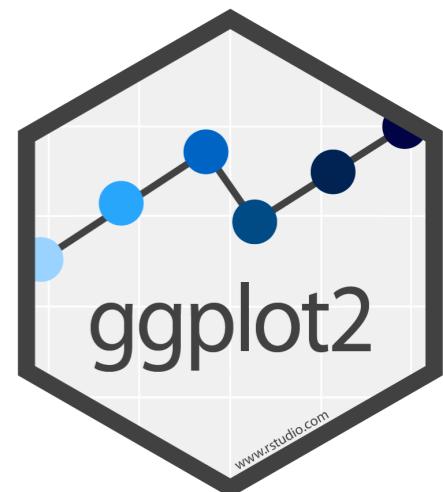
```
# dplyr
crime.by.state <- read.csv("CrimeStatebyState.csv")
final <- crime.by.state %>%
  filter(State=="New York", Year==2005) %>%
  arrange(desc(Count)) %>%
  select(Type.of.Crime, Count) %>%
  mutate(Proportion=Count/sum(Count)) %>%
  group_by(Type.of.Crime) %>%
  summarise(num.types = n(), counts = sum(Count))
```



Readability

```
# base R
crime.by.state <- read.csv("CrimeStatebyState.csv")
crime.ny.2005 <- crime.by.state[crime.by.state$Year==2005 &
                                crime.by.state$State=="New York",
                                c("Type.of.Crime", "Count")]
crime.ny.2005 <- crime.ny.2005[order(crime.ny.2005$Count,
                                         decreasing=TRUE), ]
crime.ny.2005$Proportion <- crime.ny.2005$Count /
                            sum(crime.ny.2005$Count)
summary1 <- aggregate(Count ~ Type.of.Crime,
                       data=crime.ny.2005,
                       FUN=sum)
summary2 <- aggregate(Count ~ Type.of.Crime,
                       data=crime.ny.2005,
                       FUN=length)
final <- merge(summary1, summary2,
                by="Type.of.Crime")
```

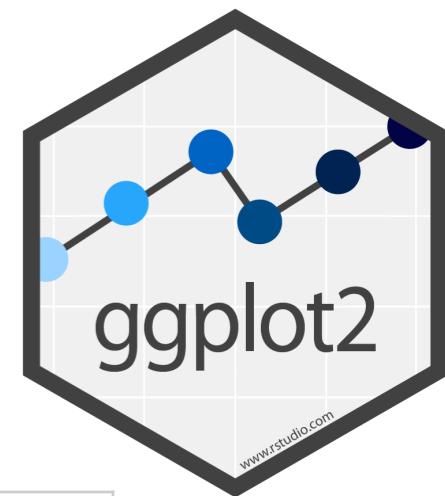
Philosophy of visualisation



"In brief, the grammar tells us that a statistical graphic is a mapping from data to aesthetic attributes (colour, shape, size) of geometric objects (points, lines, bars). The plot may also contain statistical transformations of the data and is drawn on a specific coordinates system."



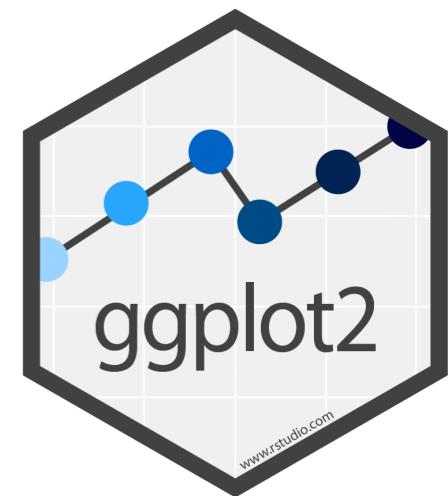
Philosophy of visualisation



1	Data		The raw data that you want to plot
2	Geometries	<code>geom_</code>	The geometric shapes that will represent the data.
3	Aethetics	<code>aes()</code>	Aesthetics of the geometric and statistical objects, such as color, size, shape and position.
4	Scales	<code>scale_</code>	Maps between the data and the aesthetic dimensions, such as data range to plot width or factor values to colors
5	Statistical transformations	<code>stat_</code>	Statistical summaries of the data that can be plotted, such as quantiles, fitted curves (loess, linear models, etc.), sums and so on.
6	Coordinate systems	<code>coord_</code>	The transformation used for mapping data coordinates into the plane of the data rectangle.
7	Facets	<code>facet_</code>	The arrangement of the data into a grid of plots (also known as latticing, trellising or creating small multiples).
8	Visual Themes	<code>theme</code>	The overall visual defaults of a plot: background, grids, axe, default typeface, sizes, colors, etc.

Hadley Wickham (2010) A Layered Grammar of Graphics, Journal of Computational and Graphical Statistics, 19:1, 3-28, DOI: 10.1198/jcgs.2009.07098

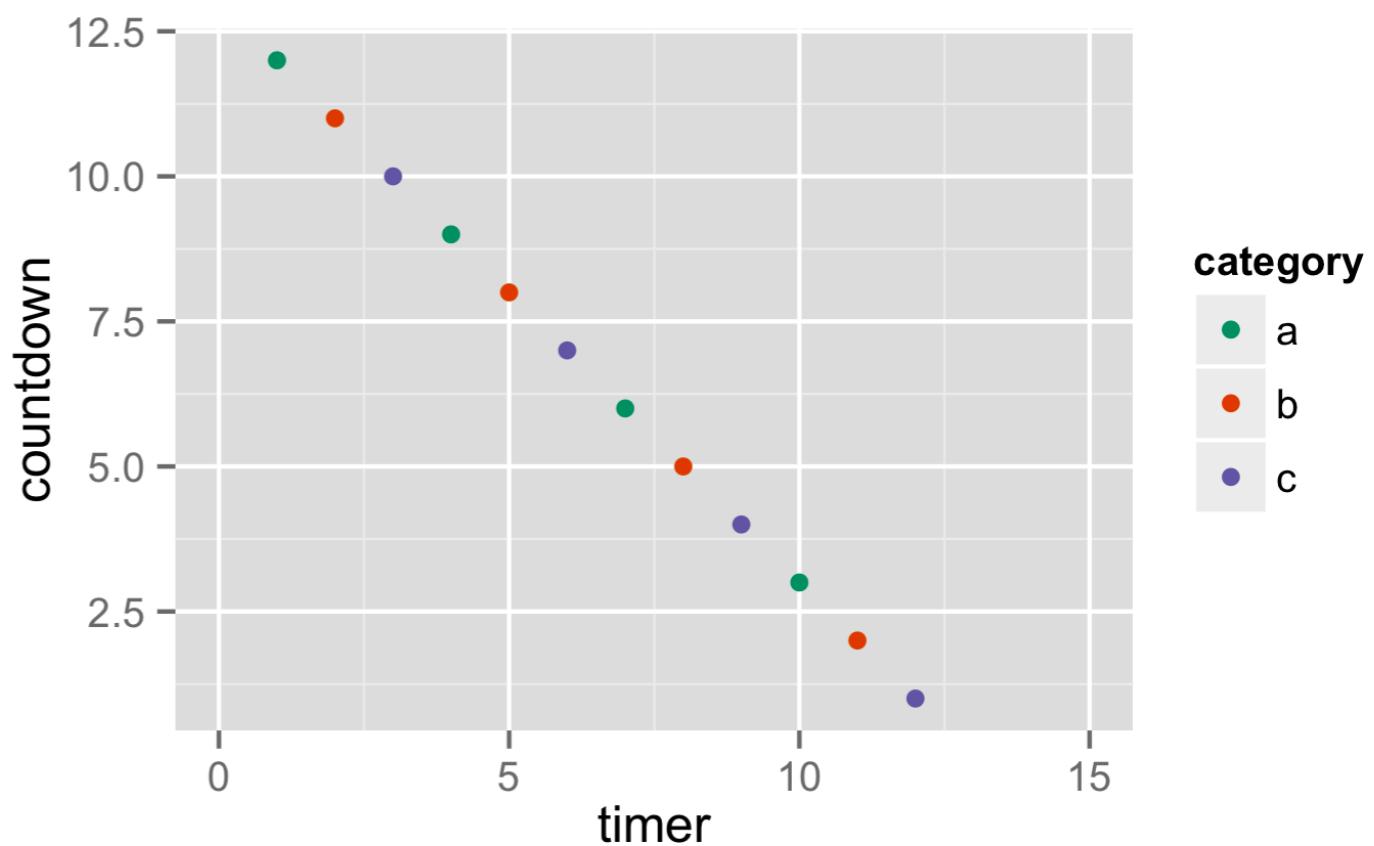
Philosophy of visualisation



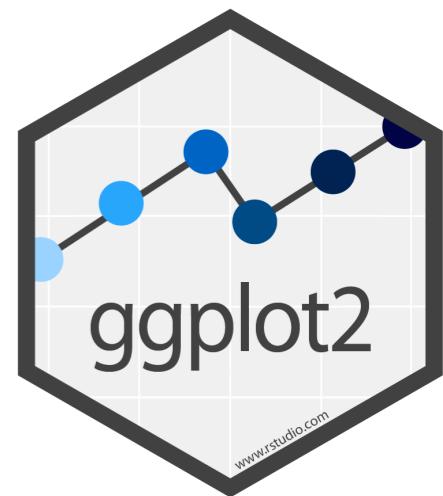
```
# Create some data for our example
some.data <- data.frame(timer = 1:12,
                         countdown = 12:1,
                         category = factor(letters[1:3]))

# Generate the plot
some.plot <- ggplot(data = some.data,
                     aes(x = timer, y = countdown)) +
  geom_point(aes(colour = category)) +
  scale_x_continuous(limits = c(0, 15)) +
  scale_colour_brewer(palette = "Dark2") +
  coord_fixed(ratio=1)

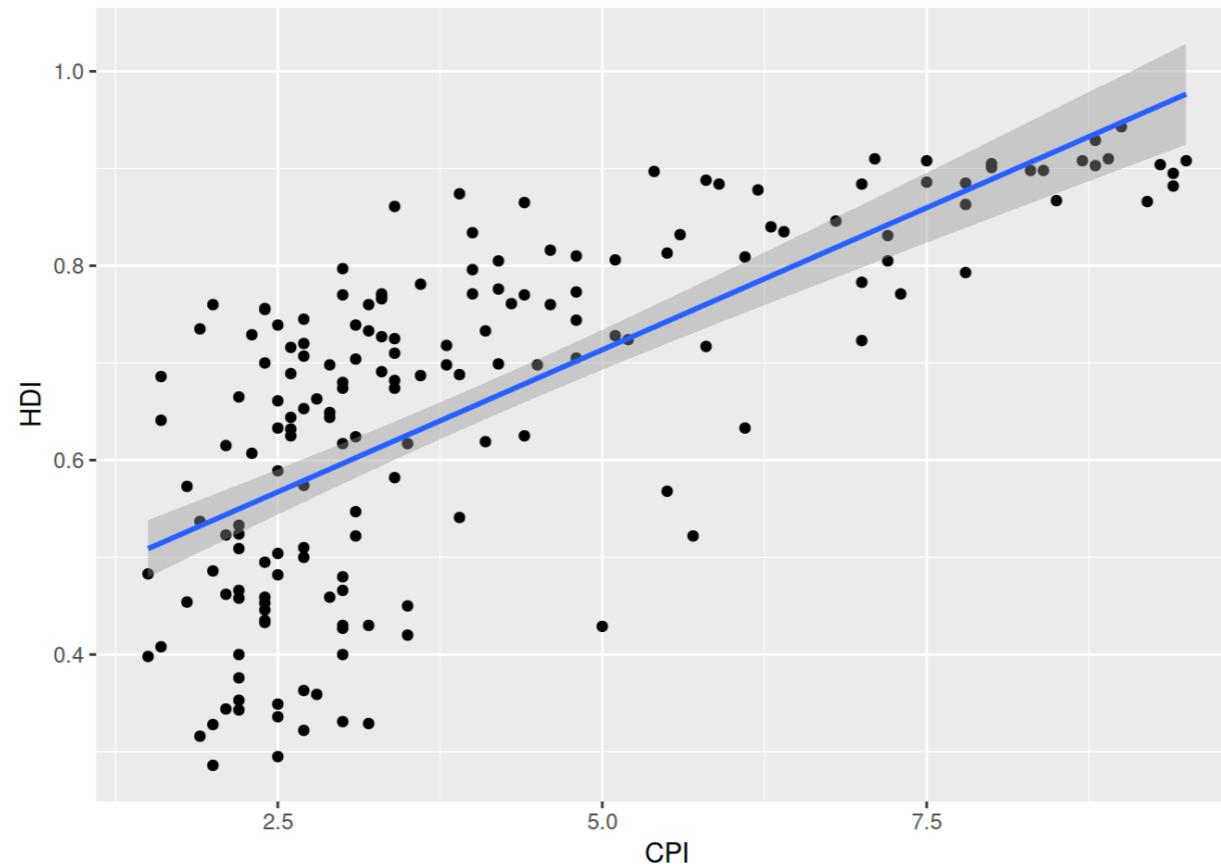
# Display the plot
some.plot
```



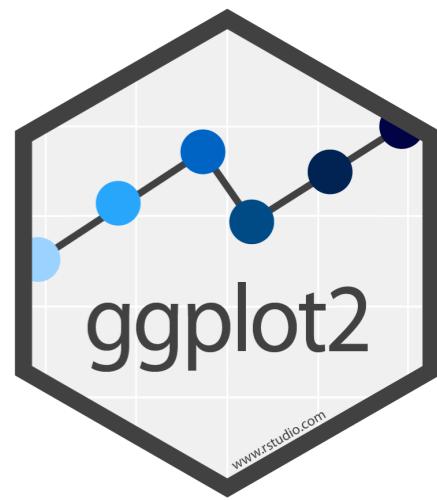
Philosophy of visualisation



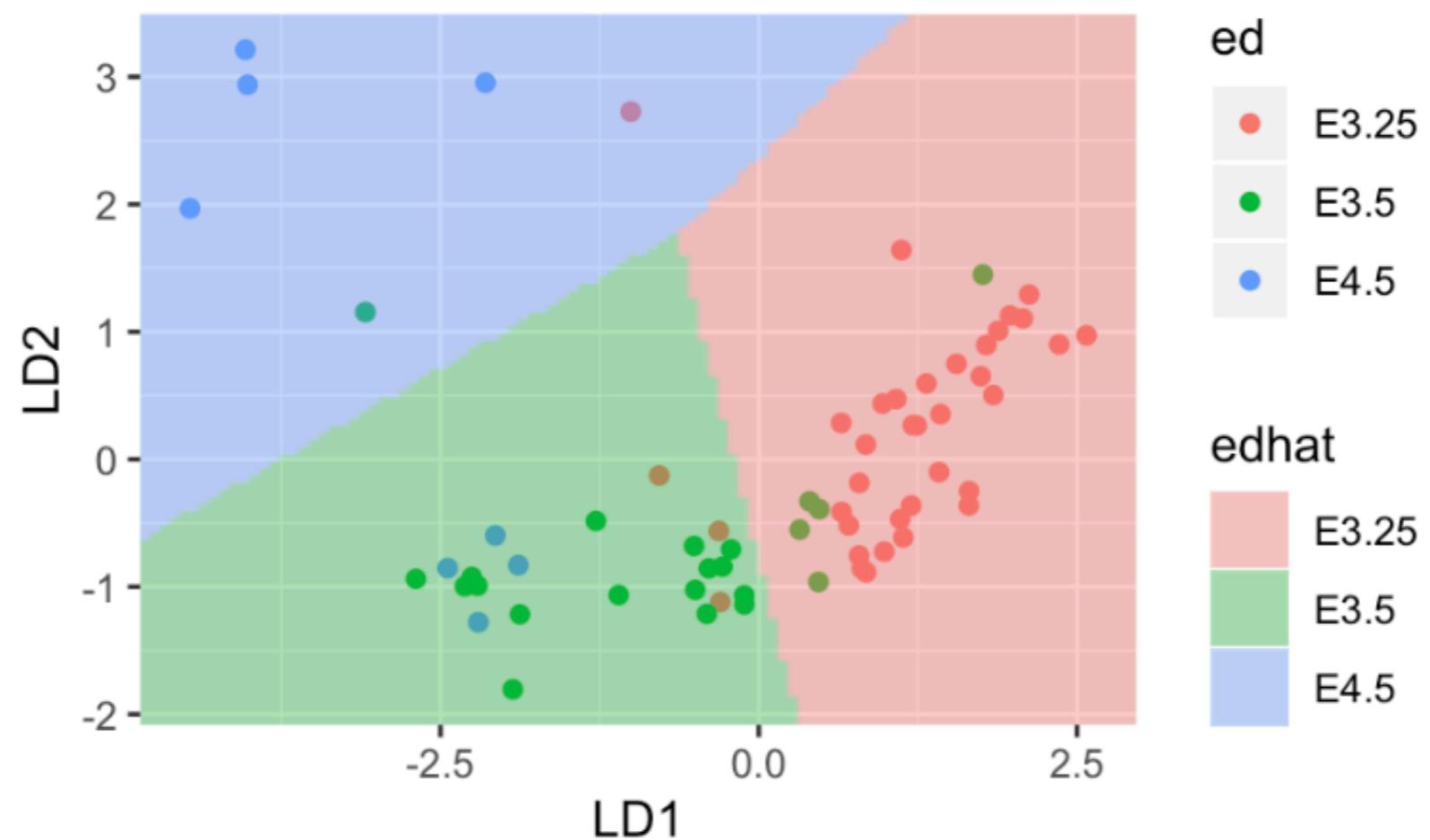
```
ggplot(dat, aes(x = CPI, y = HDI)) +  
  geom_point() +  
  geom_smooth(method = "lm")
```



Philosophy of visualisation



```
ggplot() +  
  geom_point(aes(x = LD1, y = LD2, colour = ed), data = ec_rot) +  
  geom_raster(aes(x = LD1, y = LD2, fill = edhat),  
              data = ec_grid, alpha = 0.4, interpolate = TRUE) +  
  scale_x_continuous(expand = c(0, 0)) +  
  scale_y_continuous(expand = c(0, 0)) +  
  coord_fixed()
```



Ease of assistance

The internet will make those bad words go away



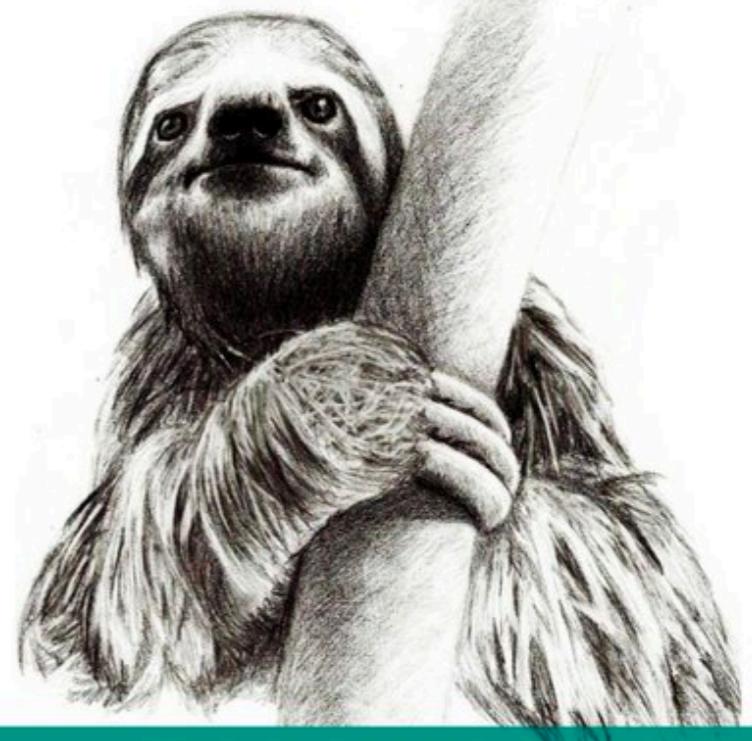
Essential

Googling the
Error Message

O RLY?

The Practical Developer
@ThePracticalDev

Cutting corners to meet arbitrary management deadlines



Essential

Copying and Pasting
from Stack Overflow

O'REILLY®

The Practical Developer
@ThePracticalDev

Base R statistical output

```
tfit <- t.test(1:10, 10:20)
tfit

## Welch Two Sample t-test

## data: 1:10 and 10:20
## t = -6.862, df = 18.998, p-value = 0.000001513
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -12.397677 -6.602323
## sample estimates:
## mean of x mean of y
##      5.5      15.0
```

Base R statistical output

```
tfit <- t.test(1:10, 10:20)
tfit

## Welch Two Sample t-test

## data: 1:10 and 10:20
## t = -6.862, df = 18.998, p-value = 0.000001513
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -12.397677 -6.602323
## sample estimates:
## mean of x mean of y
##      5.5      15.0
```

P value:
tfit\$p.value

Base R statistical output

```
lmfit <- lm(mpg ~ wt, mtcars)
lmfit

##
## Call:
## lm(formula = mpg ~ wt, data = mtcars)
##
## Coefficients:
## (Intercept)          wt
##            37.285      -5.344

summary(lmfit)

##
## Call:
## lm(formula = mpg ~ wt, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max 
## -4.5432 -2.3647 -0.1252  1.4096  6.8727 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 37.2851   1.8776  19.858 < 2e-16 ***
## wt          -5.3445   0.5591  -9.559 1.29e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.046 on 30 degrees of freedom
## Multiple R-squared:  0.7528, Adjusted R-squared:  0.7446 
## F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```

Base R statistical output

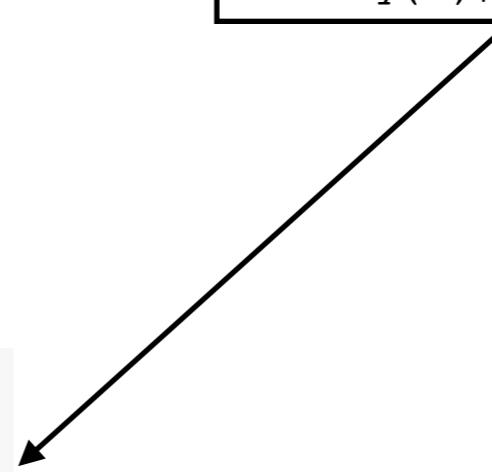
```
lmfit <- lm(mpg ~ wt, mtcars)
lmfit

## 
## Call:
## lm(formula = mpg ~ wt, data = mtcars)
##
## Coefficients:
## (Intercept)          wt
##            37.285       -5.344

summary(lmfit)

## 
## Call:
## lm(formula = mpg ~ wt, data = mtcars)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -4.5432 -2.3647 -0.1252  1.4096  6.8727
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 37.2851    1.8776  19.858 < 2e-16 ***
## wt          -5.3445    0.5591  -9.559 1.29e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.046 on 30 degrees of freedom
## Multiple R-squared:  0.7528, Adjusted R-squared:  0.7446 
## F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```

P value:
summary(x)\$coefficients[2, 4]



New packages are tidy



```
library(broom)
```

```
tidy(tfit)
# A tibble: 1 x 10
  estimate estimate1 estimate2 statistic p.value parameter conf.low conf.high method
  <dbl>     <dbl>     <dbl>     <dbl>    <dbl>     <dbl>     <dbl>     <chr>
1     -9.5      5.5       15     -6.86 1.51e-6    19.0     -12.4     -6.60 Welch...
# ... with 1 more variable: alternative <chr>
```

```
tidy(lmfit)
# A tibble: 2 x 5
  term      estimate std.error statistic p.value
  <chr>        <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)   37.3      1.88     19.9  8.24e-19
2 wt            -5.34     0.559    -9.56 1.29e-10
```

P value:
fit\$p.value

New packages are tidy



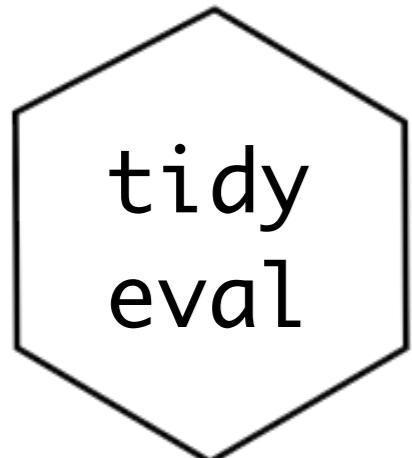
```
glance(lmfit)
## Summary statistics
# A tibble: 1 x 11
  r.squared adj.r.squared sigma statistic p.value    df logLik     AIC     BIC
* <dbl>        <dbl> <dbl>      <dbl>    <dbl> <int> <dbl> <dbl> <dbl>
1 0.753        0.745  3.05      91.4  1.29e-10     2 -80.0  166.  170.
# ... with 2 more variables: deviance <dbl>, df.residual <int>

augment(lmfit)
## Fitted values and residuals (information about the model)
# A tibble: 32 x 10
  .rownames   mpg     wt .fitted .se.fit .resid    .hat .sigma .cooksdi
* <chr>     <dbl> <dbl>    <dbl>   <dbl>    <dbl> <dbl>   <dbl>    <dbl>
1 Mazda RX4  21    2.62     23.3   0.634   -2.28  0.0433  3.07  1.33e-2
2 Mazda RX~  21    2.88     21.9   0.571   -0.920  0.0352  3.09  1.72e-3
3 Datsun 7~  22.8   2.32     24.9   0.736   -2.09  0.0584  3.07  1.54e-2
4 Hornet 4~  21.4   3.22     20.1   0.538    1.30  0.0313  3.09  3.02e-3
5 Hornet S~  18.7   3.44     18.9   0.553   -0.200  0.0329  3.10  7.60e-5
6 Valiant    18.1   3.46     18.8   0.555   -0.693  0.0332  3.10  9.21e-4
7 Duster 3~  14.3   3.57     18.2   0.573   -3.91  0.0354  3.01  3.13e-2
8 Merc 240D  24.4   3.19     20.2   0.539    4.16  0.0313  3.00  3.11e-2
9 Merc 230   22.8   3.15     20.5   0.540    2.35  0.0314  3.07  9.96e-3
10 Merc 280   19.2   3.44    18.9   0.553    0.300  0.0329  3.10  1.71e-4
# ... with 22 more rows, and 1 more variable: .std.resid <dbl>
```

New packages are tidy

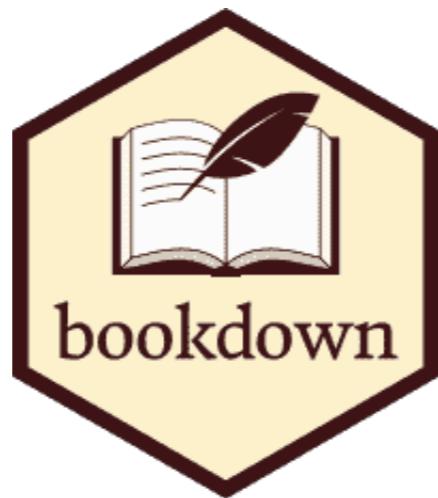


Standardizes the interface for fitting models as well as the return values



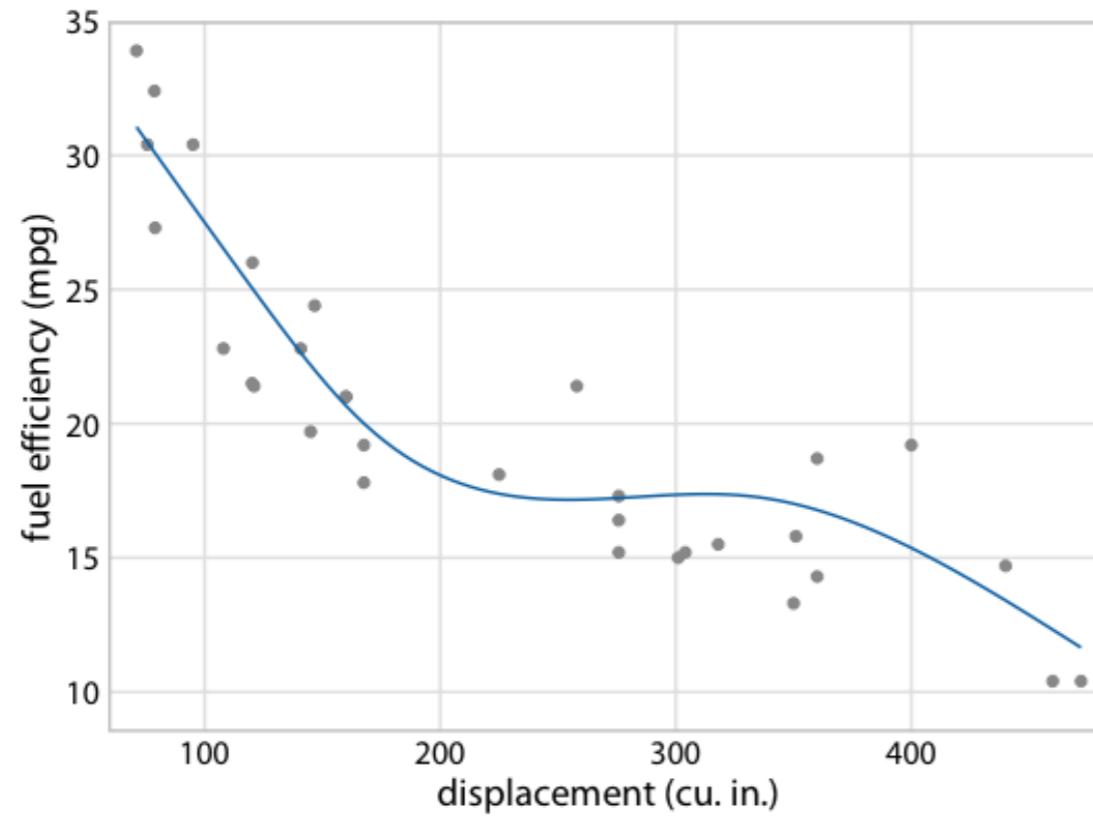
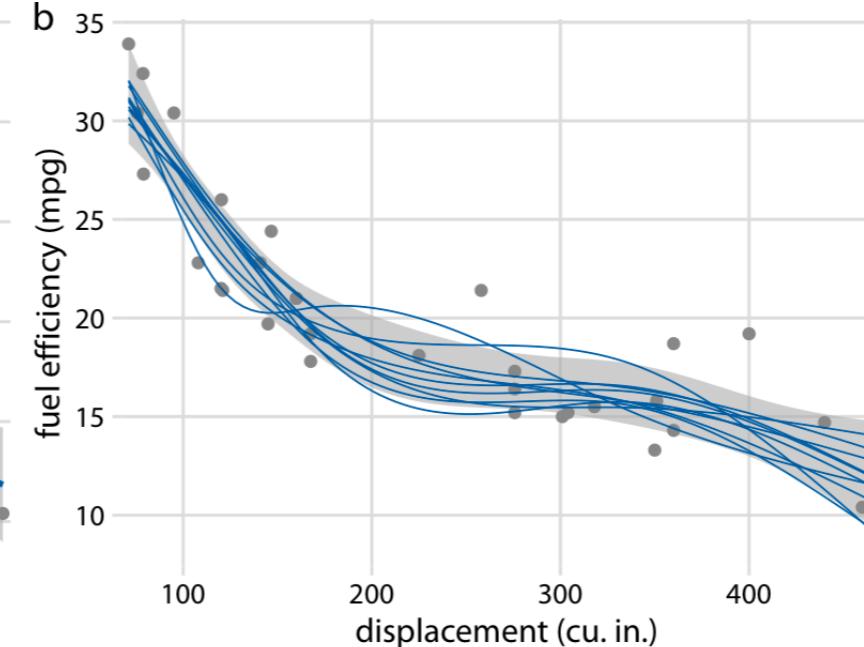
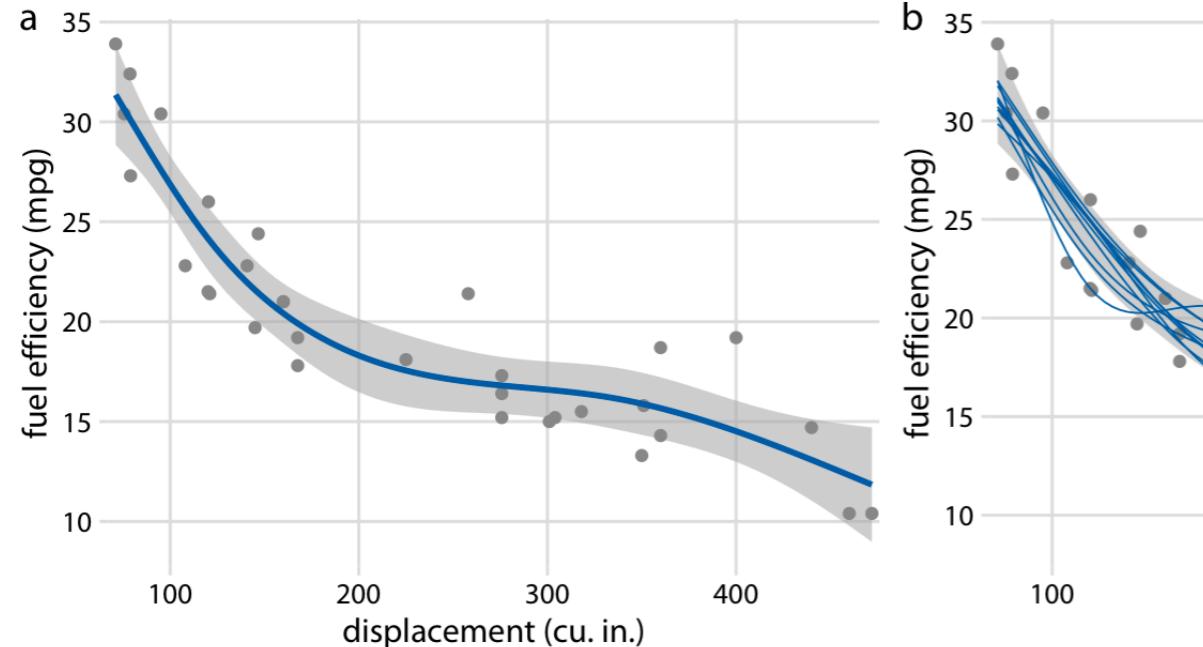
For writing functions around tidyverse pipelines and grammar

<https://tidyeval.tidyverse.org/>



(not tidy per se) Consistent documentation format

Hypothetical outcome plots



```
install.packages("gganimate")
```

Corporate R takeover

How to create BBC style graphics

- Make a line chart
- Make a multiple line chart
- Make a bar chart
- Make a stacked bar chart
- Make a grouped bar chart
- Make a dumbbell chart
- Make a histogram
- Make changes to the legend
- Make changes to the axes
- Add annotations
- Work with small multiples
- Do something else entirely

BBC Visual Journalism data team's cookbook for R graphics

Last updated: 2019-01-16

How to create BBC style graphics

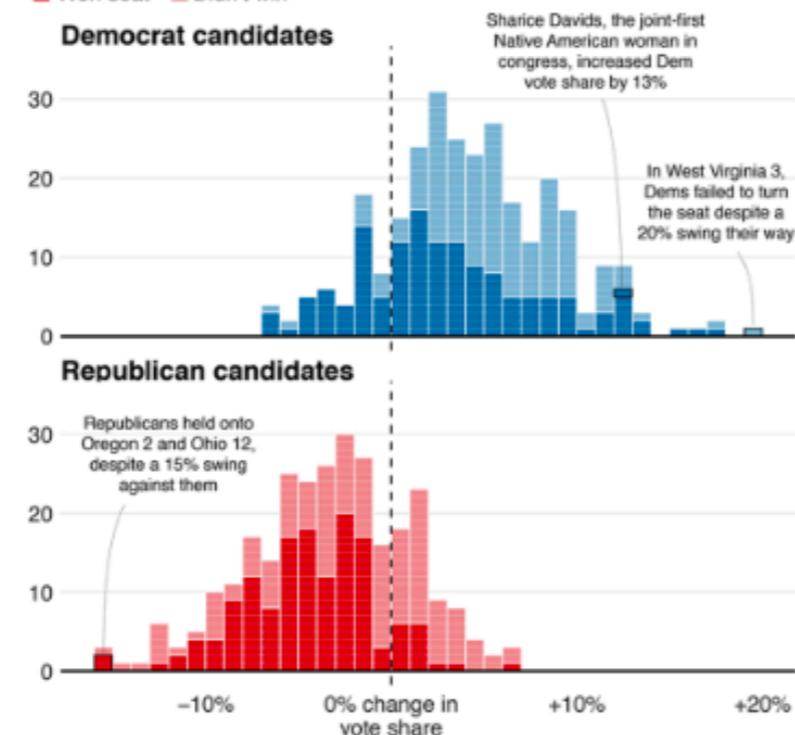
At the BBC data team, we have developed an R package and an R cookbook to make the process of creating publication-ready graphics in our in-house style using R's ggplot2 library a more reproducible process, as well as making it easier for people new to R to create graphics.

The cookbook below should hopefully help anyone who wants to make graphics like these:

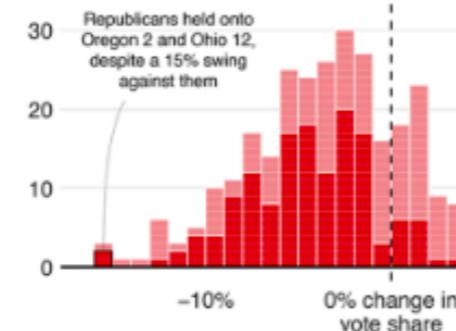
Blue wave

■ Won seat ■ Didn't win

Democrat candidates

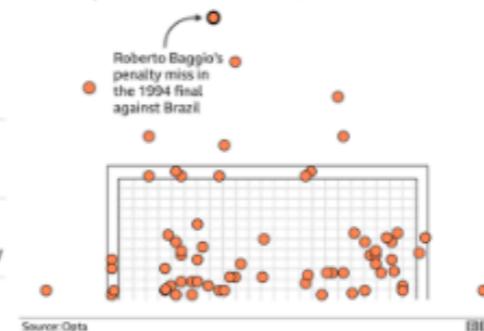


Republican candidates



Where penalties are saved

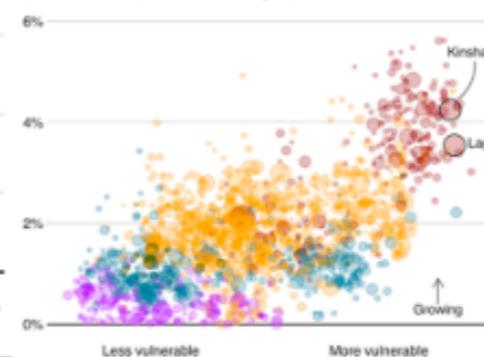
World Cup shootout misses and saves, 1982-2014



Fast-growing cities face worse climate risks

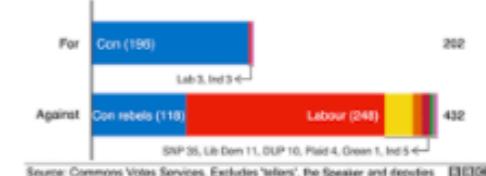
Population growth 2018-2035 over climate change vulnerability

● Africa ● Asia ● Americas ● Europe ● Oceania



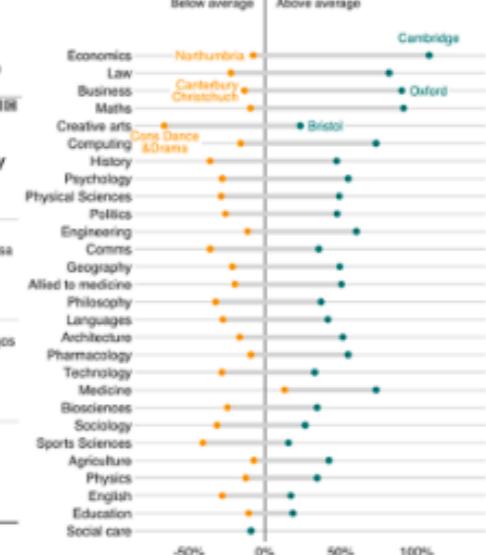
Source: Verisk Maplecroft. Circle size represents current population.

MPs rejected Theresa May's deal by 230 votes



Earnings vary across unis even within subjects

Impact on men's earnings relative to the average degree



install.packages("bbplot")

We'll get to how you can put together the various elements of these graphics, but let's get the admin out of the way first...



Top Tier Companies using R

The following is a list of top brands or large organizations using R.

1. Facebook – For behavior analysis related to status updates and profile pictures.
2. Google – For advertising effectiveness and economic forecasting.
3. Twitter – For data visualization and semantic clustering
4. Microsoft – Acquired Revolution R company and use it for a variety of purposes.
5. Uber – For statistical analysis
6. Airbnb – Scale data science.
7. IBM – Joined R Consortium Group
8. ANZ – For credit risk modeling
9. HP
10. Ford
11. Novartis
12. Roche
13. New York Times – For data visualization
14. McKinsey
15. BCG
16. Bain

Analytics and Consulting Companies using R

The below list comprises of niche analytics companies as well as consulting companies providing analytics or market research services.

1. A.T. Kearney
2. AbsolutData
3. AC Nielsen
4. Accenture
5. Bain & Company
6. Booz Allen Hamilton
7. Capgemini
8. Convergint
9. Deloitte Consulting
10. Evalueserve
11. EXL
12. EY
13. Fractal Analytics
14. Gartner
15. Genpact
16. IBM
17. KPMG
18. Latent View
19. Manthan Systems
20. McKinsey & Company
21. Mu Sigma
22. PricewaterhouseCoopers
23. SIBIA Analytics
24. Simplify360
25. SmartCube
26. Target
27. The Boston Consulting Group
28. Tiger Analytics
29. Tower Watson
30. WNS
31. ZS Associate

IT Companies using R

It includes major companies providing IT and professional services using R in India and other parts of the world.

1. Accenture
2. Amadeus IT Group
3. Capgemini
4. Cognizant
5. CSC
6. HCL Technologies
7. Hexaware Technologies
8. HP
9. IBM
10. IGATE
11. Infosys
12. Larsen & Toubro Infotech
13. Microsoft
14. Mindtree
15. Mphasis
16. NIIT Tech
17. Oracle Financial Services Software
18. Paytm
19. Snapdeal
20. R Systems Ltd
21. Tata Consultancy Services
22. Tech Mahindra
23. Wipro

Financial Institutions

It includes major US and European Banks, Insurance Companies and Other financial institutions using R.

1. American Express
2. ANZ
3. Bank of America
4. Barclays Bank
5. Bazaj allianz Insurance
6. Bharti Axa insurance
7. Blackrock
8. Citibank
9. Dun & Bradstreet
10. Fidelity
11. HSBC
12. JP Morgan
13. KeyBank
14. Lloyds Banking
15. RBS
16. Standard Chartered
17. UBS
18. Wells Fargo
19. Goldman Sachs
20. Morgan Stanley
21. PNC Bank
22. Citizens Bank
23. Fifth Third Bank



and



Vicki Boykis
@vboykis



Have been extremely curious about this for a while now, so I decided to create a poll.

"As someone titled 'data scientist' in 2019, I spend most of (60%+) my time:"

("Other") also welcome, add it in the replies.

195 10:17 AM - Jan 28, 2019



6% Picking features/models

67% Cleaning data/Moving data

4% Deploying models in prod

23% Analyzing/presenting data

2,116 votes • Final results



and



Sam Langton

@sh_langton

[Follow](#)



It's only recently become clear to me that novices can help older and more experienced **#rstats** users simply by (1) knowing **#tidyverse** fundamentals, and (2) knowing how to Google them. The beginner/expert sharing of skills can be a two-way street if the latter is open to it.

6:33 AM - 4 Feb 2019



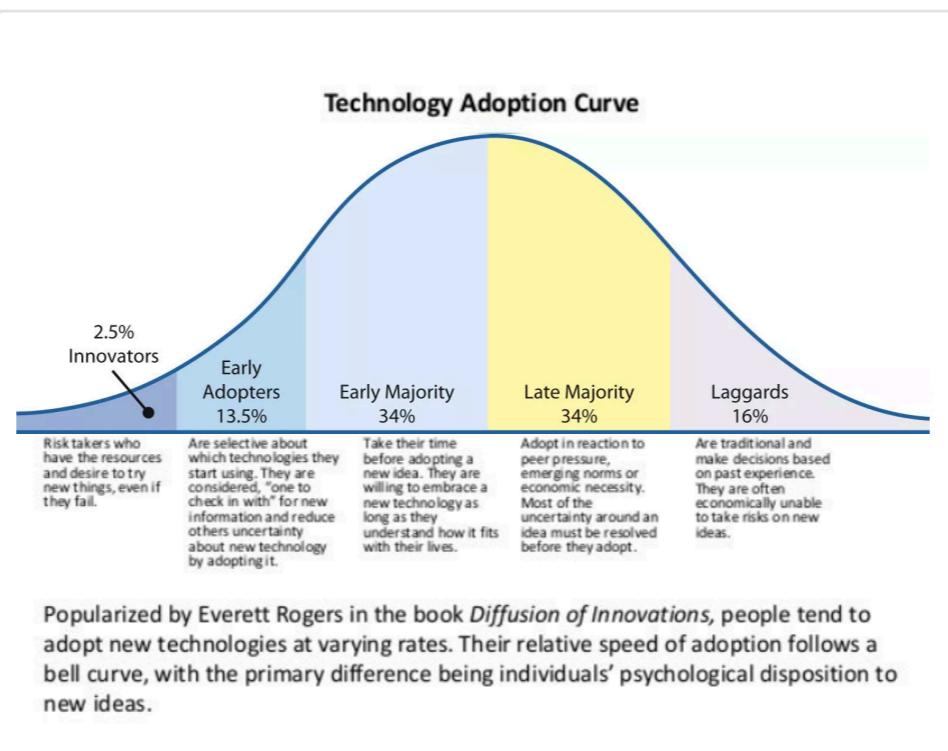
Claus Wilke

@ClausWilke

Following

Replies to @hadleywickham @thosjeeper

I think this graphic applies here. For tidyverse adoption we're maybe at the beginning of "late majority" at this time. We need to keep in mind that this is still very new technology, even if it seems so familiar to us. dplyr is ~5 years old. tidyverse less than 3.



11:18 AM - 5 Feb 2019

4 Retweets 25 Likes



2

4

25



Tweet your reply

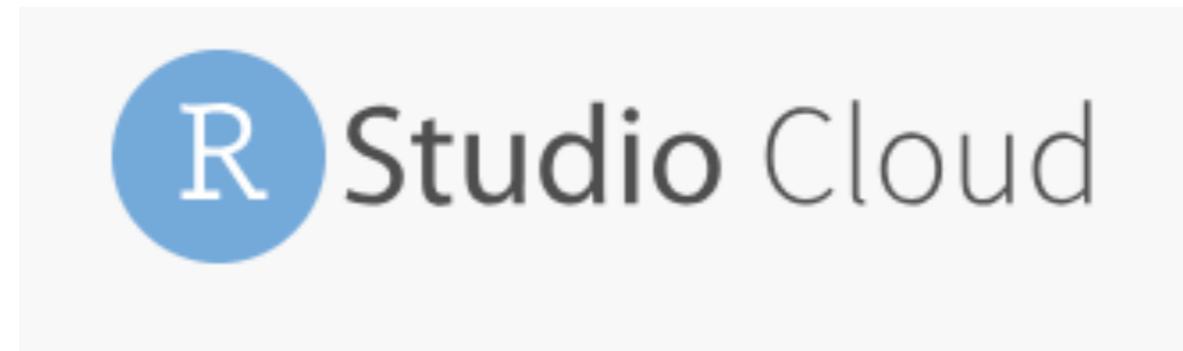


Hadley Wickham @hadleywickham · Feb 5

Replies to @ClausWilke @thosjeeper

I think we're probably still in early majority





Screenshot of the R Studio Cloud interface in a Chrome browser window. The title bar shows 'R Studio Cloud' and the URL 'https://rstudio.cloud/project/232512'. The interface includes a sidebar with 'Your Workspace / IntroStatistics', a main workspace with two R script files ('LearningToolsWeek01.R' and 'LearningToolsWeek06.R'), a console showing R code and output, and a file browser on the right.

LearningToolsWeek01.R

```
dpois(x = 3, lambda = 4.21)
0:20
expected_probability = dpois(x = 0:20, lambda = 4.21)
length(number_of_extinctions)
76 * expected_probability
expected_combined <- c(5.878568, 9.999264, 14.032300, 14.768996, 12.435494, 8.725572, 5.247808, 4.911998)
observed_combined <- c(13, 15, 16, 7, 10, 4, 2, 9)
```

LearningToolsWeek06.R

```
0:20
[1] 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
> expected_probability = dpois(x = 0:20, lambda = 4.21)
> length(number_of_extinctions)
[1] 76
> 76 * expected_probability
[1] 1.128324e+00 4.750244e+00 9.999264e+00 1.403230e+01 1.476900e+01 1.243549e+01
[7] 8.725572e+00 5.247808e+00 2.761659e+00 1.291843e+00 5.438658e-01 2.081523e-01
[13] 7.302676e-02 2.364943e-02 7.111723e-03 1.996023e-03 5.252037e-04 1.300651e-04
[19] 3.042079e-05 6.740607e-06 1.418898e-06
> expected_combined <- c(5.878568, 9.999264, 14.032300, 14.768996, 12.435494, 8.725572, 5.247808, 4.911998)
> observed_combined <- c(13, 15, 16, 7, 10, 4, 2, 9)
```

Environment

Name	Size	Modified
..	42 B	Feb 25, 2019, 9:31 AM
.RData	0 B	Feb 25, 2019, 9:31 AM
.Rhistory	256 B	Feb 25, 2019, 9:31 AM
.Rprofile	205 B	Feb 25, 2019, 9:46 AM
ABDLabs.Rproj		
Chapters		
Data		
LearningTheTools		



Feb 1 Modern Master

Aesthetics

SCIENCE DOES NOT HAVE TO BE A SHAMBLING, UGLY THING. JJ ALLAIRE AND YIHUI XIE HAVE CREATED A LOVELY TEMPLATE BASED ON THE WORK OF EDWARD TUFTE. IN THIS POST, WE EXPLORE THE INS AND OUTS OF FOLLOWING IN A MASTER'S FOOTSTEPS.



“ The commonality between science and art is in trying to see profoundly - to develop strategies of seeing and showing. ”

— EDWARD TUFTE

Edward Rolf Tufte is an American statistician and professor emeritus of political science, statistics, and computer science at Yale University. He is noted for his writings on information design and as a pioneer in the field of data visualization. If you haven't looked at any of his books,

Now that that's over with, we can tell you about the R package "tufte". It makes it incredibly easy to create beautiful RMarkdown documents. Here's an [example project](#) that we put together that demonstrates the basic features. The project is hosted on RStudio Cloud - you'll need an account to open it - but don't let that deter you. An account is free, and you'll find that it's an amazing resource for doing, sharing, teaching and learning data science.

“ The leading edge in evidence presentation is in science; the leading edge in beauty is in high art. ”

— EDWARD TUFTE

https://rstudio.cloud/project/9882

TEMPORARY

R Studio Cloud

Your Workspace / Tufte Example Robby Shaver

File Edit Code View Plots Session Build Debug Profile Tools Help

Spaces

Your Workspace

New Space

Learn

Guide

What's New

Primers

DataCamp Courses

Cheat Sheets

Feedback and Questions

Info

Terms and Conditions

System Status

Tufte1.Rmd

Knit

Insert

Run

Environment History Connections

Import Dataset

Global Environment

Environment is empty

Files Plots Packages Help Viewer

New Folder Upload Delete Rename More

Cloud > project

Name	Size	Modified
..		
skeleton.bib	603 B	Apr 5, 2018, 1:53 PM
Tufte1.html	1.1 MB	Apr 5, 2018, 1:53 PM
Tufte1.Rmd	12.5 KB	Apr 16, 2018, 1:29 PM
Tufte1_cache		
Tufte1_files		

```
17
18 ``{r setup, include=FALSE}
19 library(tufte)
20 # invalidate cache when the tufte version changes
21 knitr::opts_chunk$set(tidy = FALSE, cache.extra = packageVersion('tufte'))
22 options(htmltools.dir.version = FALSE)
23 ``
24
25 # Introduction
26
27 The Tufte handout style is a style that Edward Tufte uses in his books and handouts.
Tufte's style is known for its extensive use of sidenotes, tight integration of graphics
with text, and well-set typography. This style has been implemented in LaTeX and
HTML/CSS[See Github repositories
[tufte-latex](https://github.com/tufte-latex/tufte-latex) and
[tufte-css](https://github.com/edwardtufte/tufte-css), respectively. We have ported both
implementations into the [\*\*tufte\*\* package](https://github.com/rstudio/tufte). If you
want LaTeX/PDF output, you may use the `tufte_handout` format for handouts, and
`tufte_book` for books. For HTML output, use `tufte_html`. These formats can be either
specified in the YAML metadata at the beginning of an R Markdown document (see an example
13:26 # Tufte Example
```

R Markdown

Console Terminal R Markdown Jobs

/cloud/project/

```
R version 3.5.2 (2018-12-20) -- "Eggshell Igloo"
Copyright (C) 2018 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.
```

Tufte Example

An implementation in R Markdown

JJ Allaire and Yihui Xie

2018-04-05

Introduction

The Tufte handout style is a style that Edward Tufte uses in his books and handouts.

Tufte's style is known for its extensive use of sidenotes, tight integration of graphics with text, and well-set typography. This style has been implemented in LaTeX and HTML/CSS¹, respectively. We have ported both implementations into the [tufte package](#). If you want LaTeX/PDF output, you may use the `tufte_handout` format for handouts, and `tufte_book` for books. For HTML output, use `tufte_html`. These formats can be either specified in the YAML metadata at the beginning of an R Markdown document (see an example below), or passed to the `rmarkdown::render()` function. See Allaire et al. (2017) more information about `rmarkdown`.

```
---
```

```
title: "An Example Using the Tufte Style"
author: "John Smith"
output:
  tufte::tufte_handout: default
  tufte::tufte_html: default
---
```

There are two goals of this package:

¹ See Github repositories [tufte-latex](#) and [tufte-css](#)

Allaire, JJ, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, and Winston Chang. 2017. *Rmarkdown: Dynamic Documents for R*. <https://CRAN.R-project.org/package=rmarkdown>.



In a classroom:

No time spent on setup

Easily distribute scripts, files, assignments

Access students work easily (scripts, RMD)

Chrome File Edit View History Bookmarks People Window Help

RStudio Cloud https://rstudio.cloud/spaces/11686/project/232628

R Studio Cloud Stat2810 / Stat2810 Data Aleeza Gerstein

Spaces Your Workspace Stat2810 New Space

Learn Guide What's New Primers DataCamp Courses Cheat Sheets Feedback and Questions

Info Terms and Conditions System Status

File Edit Code View Plots Session Build Debug Profile Tools Help R 3.5.2

Console Terminal x Jobs x /cloud/project/ABDLabs/ >

Environment History Connections Import Dataset Global Environment Environment is empty

Files Plots Packages Help Viewer

New Folder Upload Delete Rename More

Cloud > project > ABDLabs > Data

	Name	Size	Modified
	antibacterial immunity.csv	628 B	Feb 25, 2019, 11:02 AM
	BatTongues.csv	364 B	Feb 25, 2019, 11:02 AM
	BeerAndMosquitoes.csv	900 B	Feb 25, 2019, 11:02 AM
	bumpus.csv	8.2 KB	Feb 25, 2019, 11:02 AM
	caffeine.csv	408 B	Feb 25, 2019, 11:02 AM
	caffeineStarbucks.csv	68 B	Feb 25, 2019, 11:02 AM
	Canadian_births.csv	197 B	Feb 25, 2019, 11:02 AM
	cardiac arrests out of hospital.csv	553 B	Feb 25, 2019, 11:02 AM
	cardiac events.csv	159 B	Feb 25, 2019, 11:02 AM
	chap02e3bGuppyFatherSonAttractive... .csv	399 B	Feb 25, 2019, 11:02 AM
	chap12e2BlackbirdTestosterone.csv	353 B	Feb 25, 2019, 11:02 AM
	child seats.csv	706 B	Feb 25, 2019, 11:02 AM
	circadian mutant health.csv	1.5 KB	Feb 25, 2019, 11:02 AM

Chrome File Edit View History Bookmarks People Window Help

RStudio Cloud https://rstudio.cloud/spaces/11686/project/232628

Stat2810 / Stat2810 Data

File Edit Code View Plots Session Build Debug Profile Tools Help

Console Terminal Jobs

/cloud/project/ABDLabs/

Environment History Connections

Import Dataset Global Environment

Environment is empty

Files Plots Packages Help Viewer

Install Update Packrat

Name	Description	Version
base	The R Base Package	3.5.2
boot	Bootstrap Functions (Originally by Angelo Canty for S)	1.3-20
class	Functions for Classification	7.3-14
cluster	"Finding Groups in Data": Cluster Analysis Extended Rousseeuw et al.	2.0.7-1
codetools	Code Analysis Tools for R	0.2-15
compiler	The R Compiler Package	3.5.2
datasets	The R Datasets Package	3.5.2
foreign	Read Data Stored by 'Minitab', 'S', 'SAS', 'SPSS', 'Stata', 'Systat', 'Weka', 'dBase', ...	0.8-71
graphics	The R Graphics Package	3.5.2
grDevices	The R Graphics Devices and Support for Colours and Fonts	3.5.2
grid	The Grid Graphics Package	3.5.2
KernSmooth	Functions for Kernel Smoothing Supporting Wand & Jones (1995)	2.23-15
lattice	Trellis Graphics for R	0.20-38
MASS	Support Functions and Datasets for Venables and Ripley's MASS	7.3-51.1
Matrix	Sparse and Dense Matrix Classes and Methods	1.2-15
methods	Formal Methods and Classes	3.5.2
mgcv	Mixed GAM Computation Vehicle with Automatic Smoothness Estimation	1.8-26
nlme	Linear and Nonlinear Mixed Effects Models	3.1-137
nnet	Feed-Forward Neural Networks and Multinomial Log-Linear Models	7.3-12
parallel	Support for Parallel computation in R	3.5.2
rpart	Recursive Partitioning and Regression Trees	4.1-13
spatial	Functions for Kriging and Point Pattern Analysis	7.3-11
splines	Regression Spline Functions and Classes	3.5.2
stats	The R Stats Package	3.5.2
stats4	Statistical Functions using S4 Classes	3.5.2

Chrome File Edit View History Bookmarks People Window Help

RStudio Cloud 48% 8:30 PM

<https://rstudio.cloud/spaces/11686/projects>

R Studio Cloud

Stat2810 Projects Members Info

Aleeza Gerstein

All Projects

New Project

Assignment 2

Delete Move

Aleeza Gerstein

Created Feb 25, 2019 7:51 PM

Assignment 1

Remove

Cara Gerstein

Created Feb 25, 2019 11:51 AM

Derived from: Assignment 1 by Aleeza Gerstein

ASSIGNMENT Assignment 1

Delete Move

Aleeza Gerstein

Created Feb 25, 2019 11:04 AM

View 1 derived project ...

Options

Search Projects

List Projects

All Shared with everyone Yours

Sort Projects

By name By date created

Capacity

This space can have up to 21 more projects.

Request More Projects

R Studio Cloud alpha

Twitter GitHub LinkedIn Facebook

© 2018 RStudio Inc.

Chrome File Edit View History Bookmarks People Window Help

RStudio Cloud https://rstudio.cloud/spaces/11686/project/232635

R Studio Cloud Stat2810 / Assignment 1 Aleeza Gerstein

Spaces Your Workspace Stat2810 New Space

Learn Guide What's New Primers DataCamp Courses Cheat Sheets Feedback and Questions

Info Terms and Conditions System Status

File Edit Code View Plots Session Build Debug Profile Tools Help R 3.5.2

Assignment 1.Rmd x Go to file/function Addins

```
1 ---  
2 title: "Assignment 1"  
3 output: html_document  
4 ---  
5  
6 ```{r setup, include=FALSE}  
7 knitr::opts_chunk$set(echo = TRUE)  
8 ```  
9  
10 ## STAT 2810 - Assignment #1  
11  
12 ##### Introduction  
13  
14 Researchers have collected data on world record times in the 100 meter freestyle for both men  
and women since 1905. Our task is to analyze that data.  
15  
16 ```{r, message=FALSE}  
17:1 C Chunk 2
```

Console Terminal x R Markdown x Jobs x /cloud/project/

Environment History Connections Import Dataset Global Environment

Environment is empty

Files Plots Packages Help Viewer

New Folder Upload Delete Rename More

Name	Size	Modified
.Rhistory	0 B	Feb 25, 2019, 8:03 PM
Assignment 1.Rmd	1.3 KB	Feb 25, 2019, 8:03 PM
project.Rproj	205 B	Feb 25, 2019, 8:03 PM
Assignment_1.html	616.3 KB	Feb 25, 2019, 8:03 PM

Chrome File Edit View History Bookmarks People Window Help

RStudio Cloud https://rstudio.cloud/spaces/11686/project/232635

R Studio Cloud Stat2810 / Assignment 1

File Edit Code View Plots Session Build Debug Profile

Assignment 1.Rmd x

```
1 ---  
2 title: "Assignment 1"  
3 output: html_document  
4 ---  
5  
6 ```{r setup, include=FALSE}  
7 knitr::opts_chunk$set(echo = TRUE)  
8  
9  
10 ## STAT 2810 - Assignment #1  
11  
12 #### Introduction  
13  
14 Researchers have collected data on world record times in  
and women since 1905. Our task is to analyze that data.  
15  
16 ```{r, message=FALSE}  
17:1 C Chunk 2
```

Console Terminal x R Markdown x Jobs x

/cloud/project/

Assignment 1.html Open in Browser Find

Publish

Assignment 1

STAT 2810 - Assignment #1

Introduction

Researchers have collected data on world record times in the 100 meter freestyle for both men and women since 1905. Our task is to analyze that data.

```
require(mosaic)  
data(SwimRecords)  
head(SwimRecords)
```

```
##   year time sex  
## 1 1905 65.8 M  
## 2 1908 65.6 M  
## 3 1910 62.8 M  
## 4 1912 61.6 M  
## 5 1918 61.4 M  
## 6 1920 60.4 M
```

Exercise 1.1

Create a scatterplot for the world record time as a function of the year in which the record was set. Use separate colors for men and women, create a legend, and label your axes.

Exercise 1.2

Fit the linear model:

$$\widehat{\text{time}} = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{year} + \epsilon$$

Interpret the coefficient for *year*.

Exercise 1.3

Illustrate your model by adding it to a scatterplot.

SOLUTION:

Chrome File Edit View History Bookmarks People Window Help

RStudio Cloud 48% 8:30 PM

<https://rstudio.cloud/spaces/11686/projects>

R Studio Cloud

Stat2810 Projects Members Info

Aleeza Gerstein

All Projects

New Project

Assignment 2

Aleeza Gerstein

Created Feb 25, 2019 7:51 PM

Delete Move

Assignment 1

Cara Gerstein

Created Feb 25, 2019 11:51 AM

Derived from: Assignment 1 by Aleeza Gerstein

Remove

Assignment 1

Aleeza Gerstein

Created Feb 25, 2019 11:04 AM

View 1 derived project ...

Delete Move

Options

Search Projects

List Projects

All Shared with everyone Yours

Sort Projects

By name By date created

Capacity

This space can have up to 21 more projects.

Request More Projects

R Studio Cloud alpha

Twitter GitHub LinkedIn Facebook

© 2018 RStudio Inc.

RStudio Cloud x +

https://rstudio.cloud/learn/guide

R Studio Cloud Learn Guide What's New Primers DataCamp Courses Cheat Sheets AG Aleeza Gerstein

Spaces Your Workspace Stat2810 New Space

Get started with RStudio Cloud.

Projects

A project is the fundamental unit of work on RStudio Cloud. It encapsulates your R code, packages and data files and provides isolation from other analyses. If you are familiar with projects in the desktop RStudio IDE, an RStudio Cloud project is the same thing, plus some additional metadata for access and sharing.

Your Workspace Projects Info

To create a new project from scratch, simply press the New Project button from the Projects area. Your new project will open in the RStudio IDE.

New Project

To create a new project from an existing git repository, press the down arrow on the right side of the New Project button, and choose 'New Project from Git Repo' from the menu that appears. Note that your git credentials need to be entered each time you create a new project and are only cached for 15 minutes by default. See [Working with Git](#) below for more info on working with git.

New Project

RStudio Cloud

<https://rstudio.cloud/learn/cheat-sheets>

Learn Guide What's New Primers DataCamp Courses Cheat Sheets

Aleeza Gerstein

Work with Strings

The string package provides an easy to use toolkit for working with strings, i.e. character data, in R. This cheatsheet guides you through stringr's functions for manipulating strings. The back page provides a concise reference to regular expressions, a mini-language for describing, finding, and matching patterns in strings.

Updated October 2017

[Download](#)

Apply Functions

The purrr package makes it easy to work with lists and functions. This cheatsheet will remind you how to manipulate lists with purrr as well as how to apply functions iteratively to each element of a list or vector. The back of the cheatsheet explains how to work with list-columns. With list columns, you can use a simple data frame to organize any collection of objects in R.

Updated September 2017

[Download](#)

Data Import

The Data Import cheat sheet reminds you how to read in flat files with <http://readr.tidyverse.org/>, work with the results as tibbles, and reshape messy data with tidyr. Use tidyr to reshape your tables into tidy data, the data format that works the most seamlessly with R and the tidyverse.

Updated January 2017

[Download](#)

Data Transformation

dplyr provides a grammar for manipulating tables in R. This cheat sheet will guide you through the grammar, reminding you how to select, filter, arrange, mutate, summarise, group, and join data frames and tibbles.

Updated January 2017

[Download](#)

Sparklyr

Sparklyr provides an R interface to Apache Spark, a fast and general engine for

R Markdown

R Markdown is an authoring format that makes it easy to write reusable reports with

R Studio Cloud x

Learn Guide What's New Primers DataCamp Courses Cheat Sheets AG Aleeza Gerstein

R Studio Primers

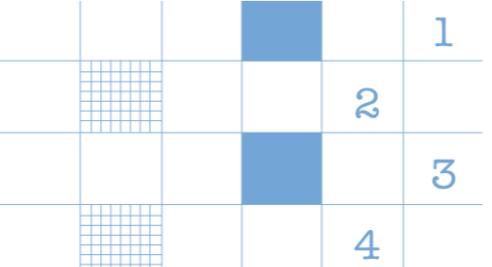
Learn data science basics with the interactive tutorials below.

The Basics



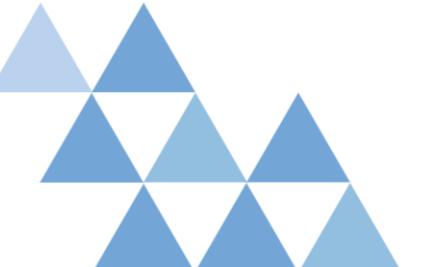
Start here to learn the skills that you will rely on in every analysis (and every primer that follows): how to inspect, visualize, subset, and transform your data, as well as how to run code.

Work with Data



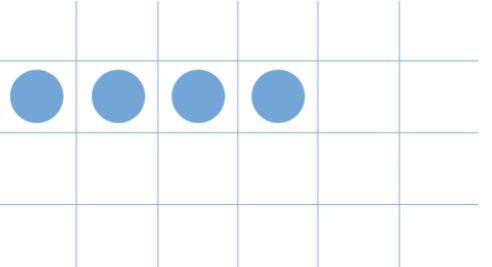
Learn the most important data handling skills in R: how to extract values from a table, subset tables, calculate summary statistics, and derive new variables.

Visualize Data



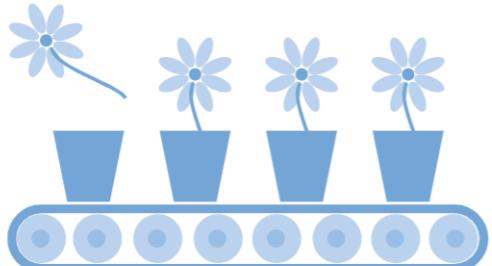
Learn how to use ggplot2 to make any type of plot with your data. Then learn the best ways to visualize patterns within values and relationships between variables.

Tidy Your Data



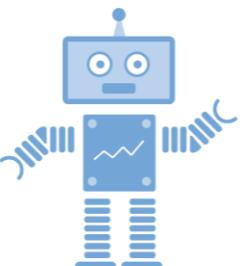
Unlock the tidyverse by learning how to make and use tidy data, the data format designed for R.

Iterate



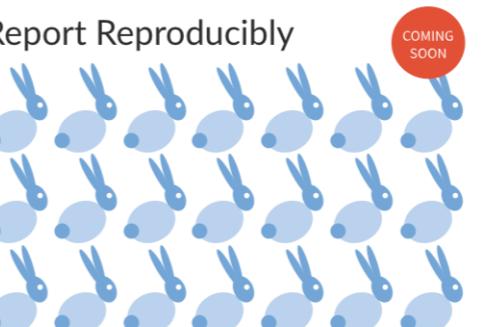
Master a core programming paradigm with the purrr package: for each ___ do ___.

Write Functions



Functions are the key to programming in R. This primer will teach you how to write and use your own reusable functions.

Report Reproducibly



Learn to report, reproduce, and parameterize your work with the best authoring format for Data Science: R Markdown.

Build Interactive Web Apps



Say hello to Shiny, R's package for building interactive web apps. Learn to turn your analyses into elegant tools to share with others.

COMING SOON

Chrome File Edit View History Bookmarks People Window Help

RStudio Cloud

https://rstudio.cloud/learn/data-camp-courses

Learn Guide What's New Primers DataCamp Courses Cheat Sheets Aleeza Gerstein

Spaces Your Workspace Stat2810 New Space

Learn Guide What's New Primers DataCamp Courses Cheat Sheets Feedback and Questions

Info Terms and Conditions System Status

DataCamp Courses

DataCamp offers more than 60 courses in R. Click any course below to learn more.

Introduction to R

Master the basics of data analysis by manipulating common data structures such as vectors, matrices and data frames.

4 hours



JONATHAN CORNELISSEN
Co-founder and CEO of DataCamp

Data Analysis in R, the data.table Way

Master core concepts in data manipulation such as subsetting, updating, indexing and joining your data using data.table.

4 hours



MATT DOWLE
Author of data.table

Data Manipulation in R with dplyr

Master techniques for data manipulation using the select, mutate, filter, arrange, and summarise functions in dplyr.

4 hours



GARRETT GROLEMUND
Data Scientist at RStudio

Data Visualization in R with ggvis

Learn to create interactive graphs to display distributions, relationships, model fits, and more using ggvis.

4 hours



GARRETT GROLEMUND
Data Scientist at RStudio

Reporting with R Markdown

Learn to create interactive analyses and automated reports with R Markdown.

3 hours



GARRETT GROLEMUND
Data Scientist at RStudio

Intermediate R

Continue your journey to become an R ninja by learning about conditional statements, loops, and vector functions.

6 hours



FILIP SCHOUWENAARS
Data Science Instructor at DataCamp

DataCamp for Stat7350

Chrome File Edit View History Bookmarks People Window Help

STAT7350- Statistical Analysis X +

https://www.datacamp.com/enterprise/stat7350-statistical-analysis-and-visualization-of-biological-data-in-r

STAT7350- Statistical Analysis and Visualization of Biological Data in R

Invite Members

DASHBOARD

MEMBERS

CUSTOM TRACKS

ASSIGNMENTS

REPORTING

TEAMS

SETTINGS

HELP

ACADEMIC

Dashboard

30% INVITED

30% ENROLLED

33% MONTHLY ACTIVE

7 Open Invitations

3 Members Enrolled

Set Assignments

Invite Members

Quickly invite Members by sharing the secure Invite Link below. You can also send individual invites by email on the [Members](#) page.

Invite Link for [myumanitoba.ca](https://www.datacamp.com/groups/shared_links/3)

https://www.datacamp.com/groups/shared_links/3 [Copy](#)

Create a Team

Sort Members into [Teams](#) based on department, skill level, and more. This makes it easier to create Assignments and view Reports for specific groups within your organization.

[Create Your First Team](#)

Did you know?

This page will continuously evolve and show more information as your organization grows.

Best Practices Guide

Checkout our Best Practices Guide to make sure you are getting the most out of your new Subscription.

[View Guide](#)

DataCamp for Stat7350

Chrome File Edit View History Bookmarks People Window Help

Organization Assignments https://www.datacamp.com/enterprise/stat7350-statistical-analysis-and-visualization-of-biological-data-in-r/assignments

STAT7350- Statistical Analysis and Visualization of Biological Data in R

Invite Members

DASHBOARD MEMBERS CUSTOM TRACKS

ASSIGNMENTS REPORTING TEAMS SETTINGS HELP

ACADEMIC

Assignments

ACTIVE ASSIGNMENTS ARCHIVED ASSIGNMENTS

NAME	ASSIGNED TO	ASSIGNED AT	DUEDATE	TYPE	COMPLETED	PENDING	OVERDUE	MORE
Basic workflow	Organization	Feb 21, 2019	Mar 1, 2019, 16:20 CST	Complete Chapter	0	0	0	>

DataCamp for Stat7350

Chrome File Edit View History Bookmarks People Window Help

Assignment

https://www.datacamp.com/enterprise/stat7350-statistical-analysis-and-visualization-of-biological-data-in-r/assignments/45469

STAT7350- Statistical Analysis and Visualization of Biological Data in R

Invite Members

DASHBOARD

MEMBERS

CUSTOM TRACKS

ASSIGNMENTS

REPORTING

TEAMS

SETTINGS

HELP

ACADEMIC

Back to assignments

Basic workflow

Organization members must complete this Chapter before Mar 1, 2019 at 16:20

Created on Feb 21, 2019

Delete Archive Edit

Search members...

0 Completed 0 Late 0 Missed

NAME	EMAIL	STATUS	COMPLETED ON
Aleeza Gerstein	gerstein@zoology.ubc.ca	In progress	Not yet completed
Jingyu Wang	wangj321@myumanitoba.ca	In progress	Not yet completed
Scott White	umwhit35@myumanitoba.ca	In progress	Not yet completed

DataCamp for Stat7350

Chrome File Edit View History Bookmarks People Window Help

Assignment How can I tell what's going to b

https://campus.datacamp.com/courses/introduction-to-git-for-data-science/basic-workflow?ex=7

DataCamp

Exercise Course Outline Terminal

How can I tell what's going to be committed?

To compare the state of your files with those in the staging area, you can use `git diff -r HEAD`. The `-r` flag means "compare to a particular revision", and `HEAD` is a shortcut meaning "the most recent commit".

You can restrict the results to a single file or directory using `git diff -r HEAD path/to/file`, where the path to the file is relative to where you are (for example, the path from the root directory of the repository).

We will explore other uses of `-r` and `HEAD` in the next chapter.

Instructions 2/3 30XP

✓ You have been put in the `dental` repository, where `data/northern.csv` has been added to the staging area. Use `git diff` with `-r` and an argument to see how files differ from the last saved revision.

2 Use a single Git command to view the changes in the file that has been staged (and *only* that file).

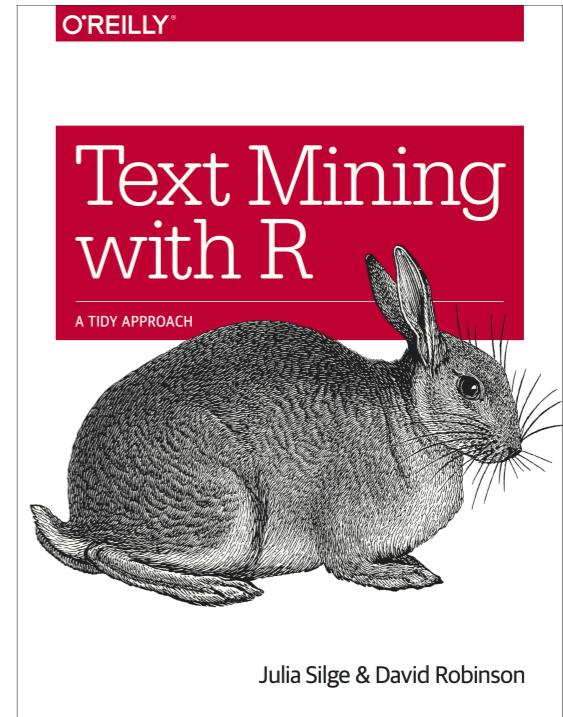
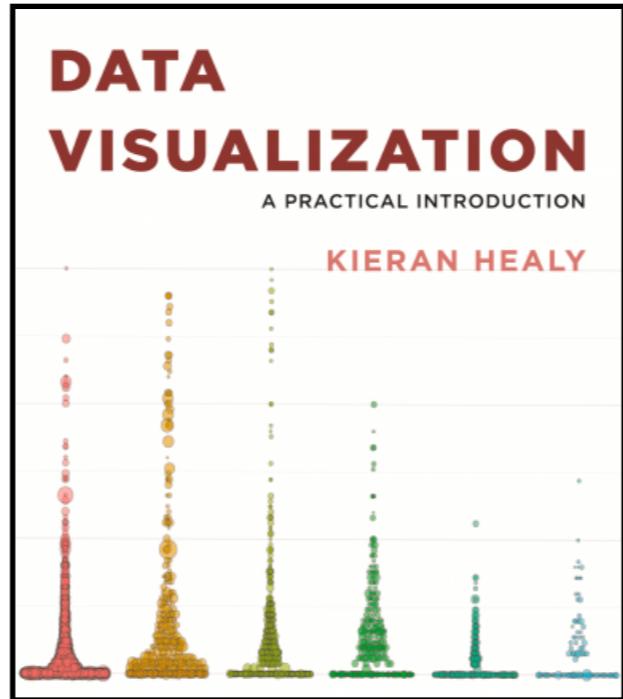
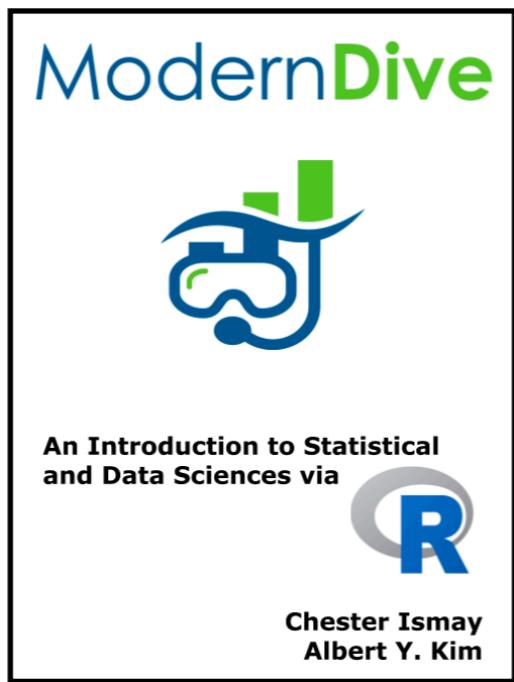
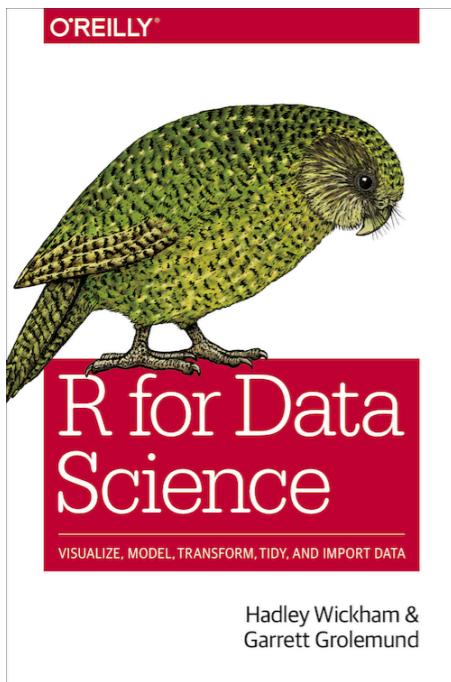
Take Hint (-9 XP)

3 `data/eastern.csv` hasn't been added to the staging area yet. Use a Git command to do this now.

```
$ cd dental
$ git add data/northern.csv
$ git diff -r
fatal: ambiguous argument 'r': unknown revision or path not in the working tree.
Use '--' to separate paths from revisions, like this:
'git <command> [<revision>...] -- [<file>...]'
$ git diff -r HEAD
fatal: ambiguous argument 'r': unknown revision or path not in the working tree.
Use '--' to separate paths from revisions, like this:
'git <command> [<revision>...] -- [<file>...]'
$ git diff -r HEAD
diff --git a/data/eastern.csv b/data/eastern.csv
index b3c1688..85053c3 100644
--- a/data/eastern.csv
+++ b/data/eastern.csv
@@ -23,3 +23,4 @@ Date,Tooth
2017-08-02,canine
2017-08-03,bicuspid
2017-08-04,canine
+2017-11-02,molar
diff --git a/data/northern.csv b/data/northern.csv
index 5eb7a96..5a2a259 100644
--- a/data/northern.csv
+++ b/data/northern.csv
@@ -22,3 +22,4 @@ Date,Tooth
2017-08-13,incisor
2017-08-13,wisdom
2017-09-07,molar
+2017-11-01,bicuspid
$ 
```

(a sampling of)

RRResources

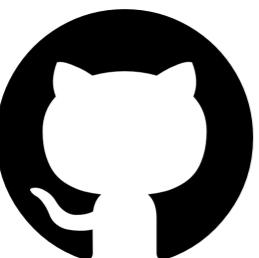


What They Forgot to Teach You

Happy gitR



Advanced Statistical Computing



R Courses

Tidy Tuesday

