Introduction to R

Gemma Dawson and Luis de Sousa

2018/10/15 (updated: 2018-10-23)

Who are we?

Gemma Dawson

www.icepack.ai
@GemmaDawson
@gemmadawsonza

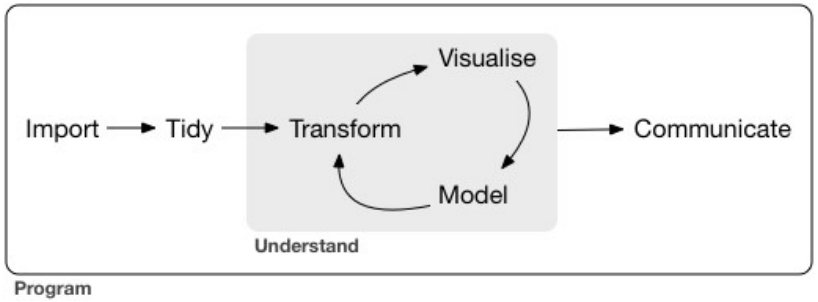
Luis de Sousa

@luisdza
@luis_de_sousa
grownconsultation.com

Who are you?


NONE
I have no idea what you're talking about!?!
NOVICE
I have opened the programme and messed around a bit.
I have successfully completed one small task.
BASICS
I do simple tasks and think I can complete a new challenge but only with Google's help!
INTERMEDIATE
I have refactored some of my old work as I have learnt new techniques.
I have completed a training course (MOOC or in person).
ADVANCED
I have completed several projects using this programme.
I have contributed to the a blog/community/forum/user group.
PROFICIENT
I have developed a package/module to help with a collaborative solution to a common problem.

Table of Contents

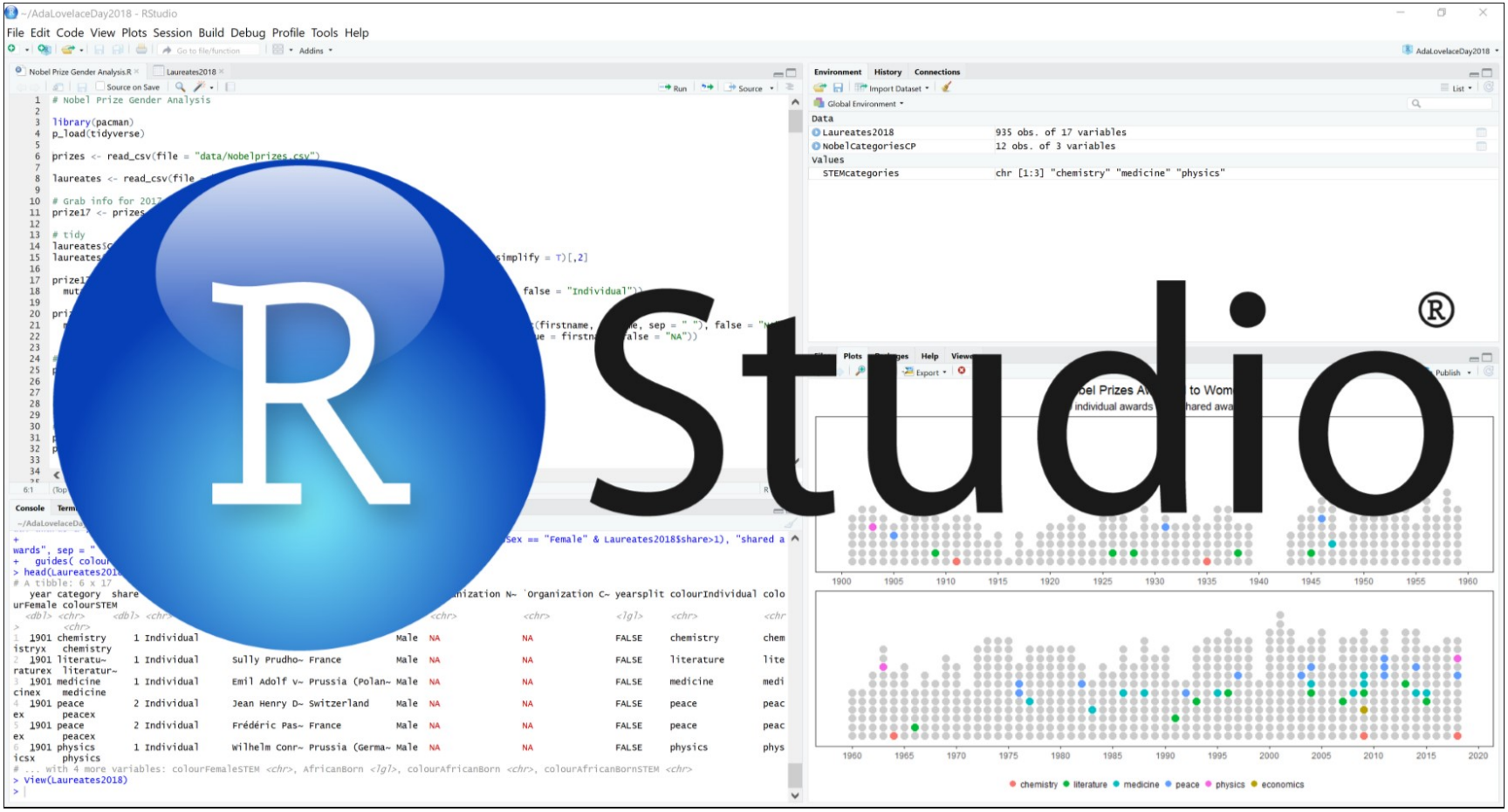


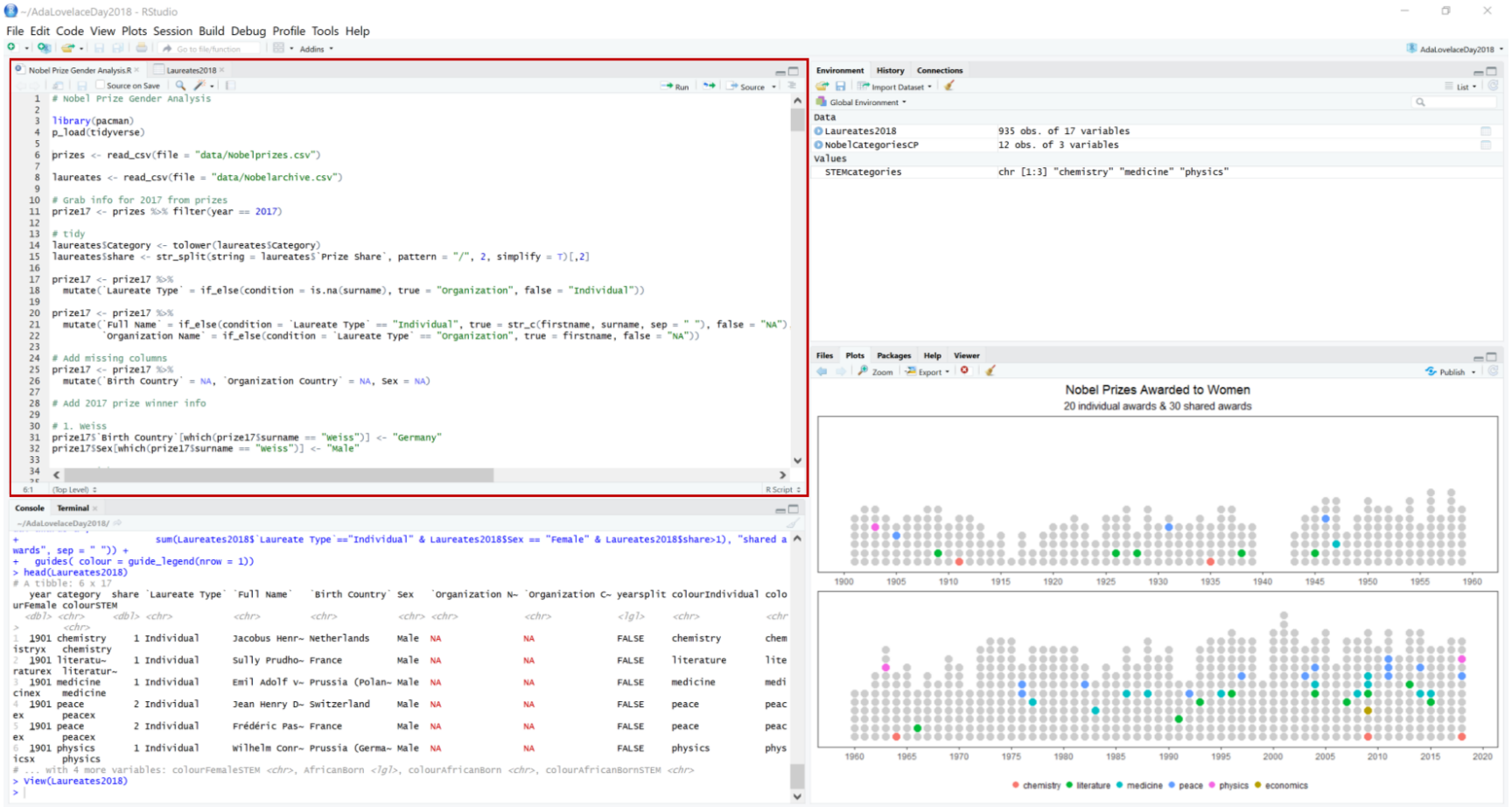
What is R?

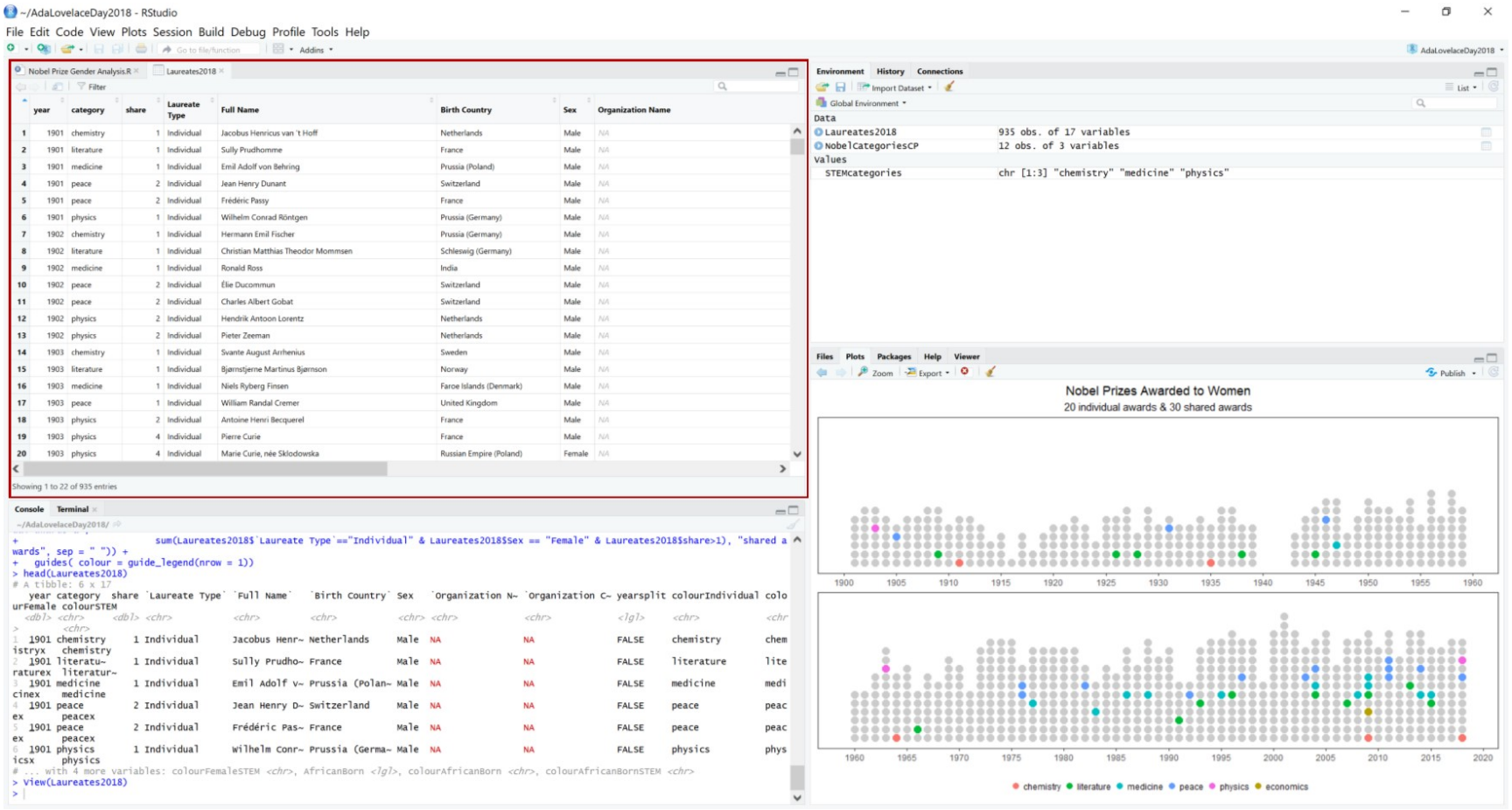
programming language and environment for statistical computing and graphics

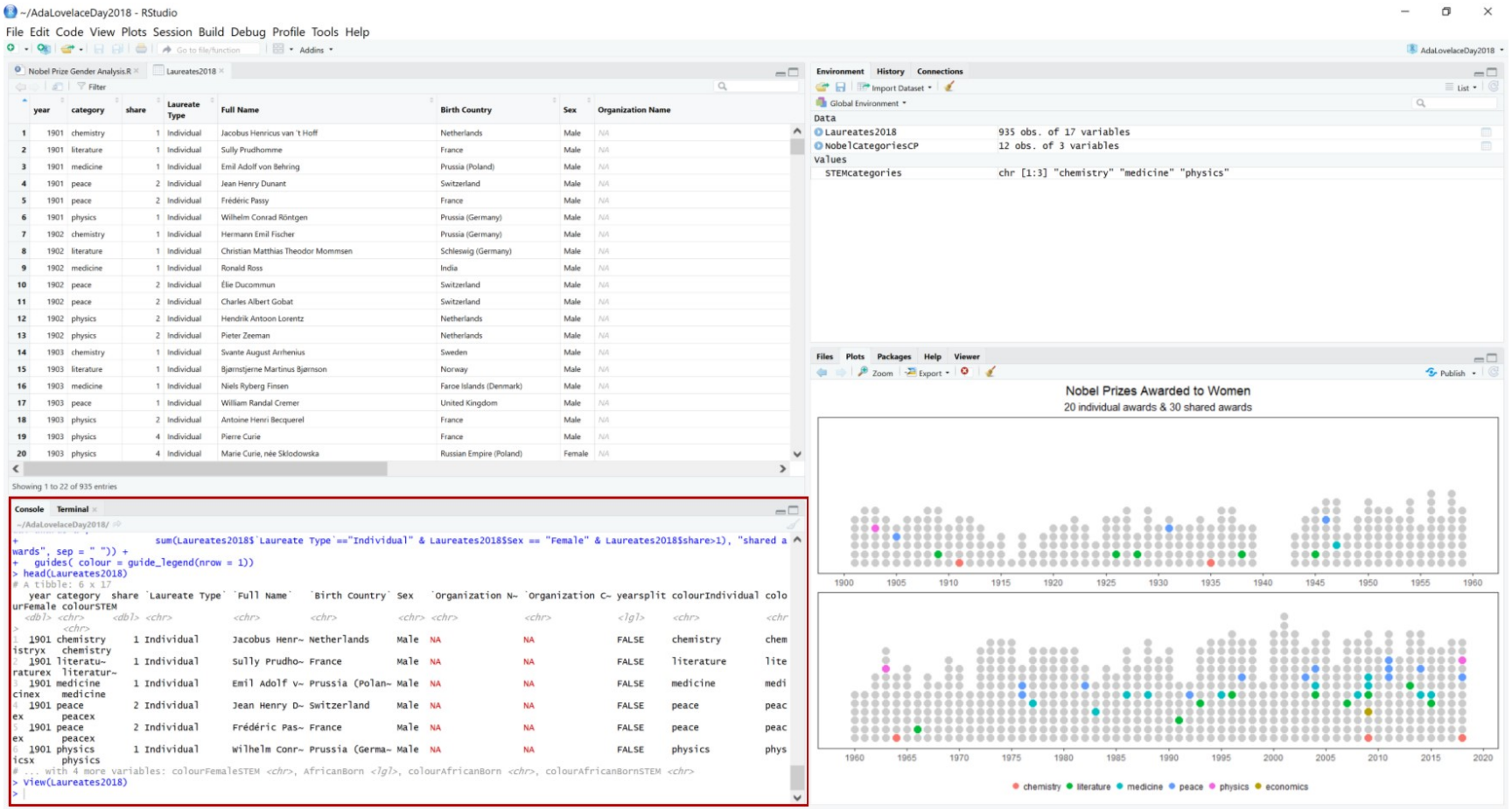
R versus Python

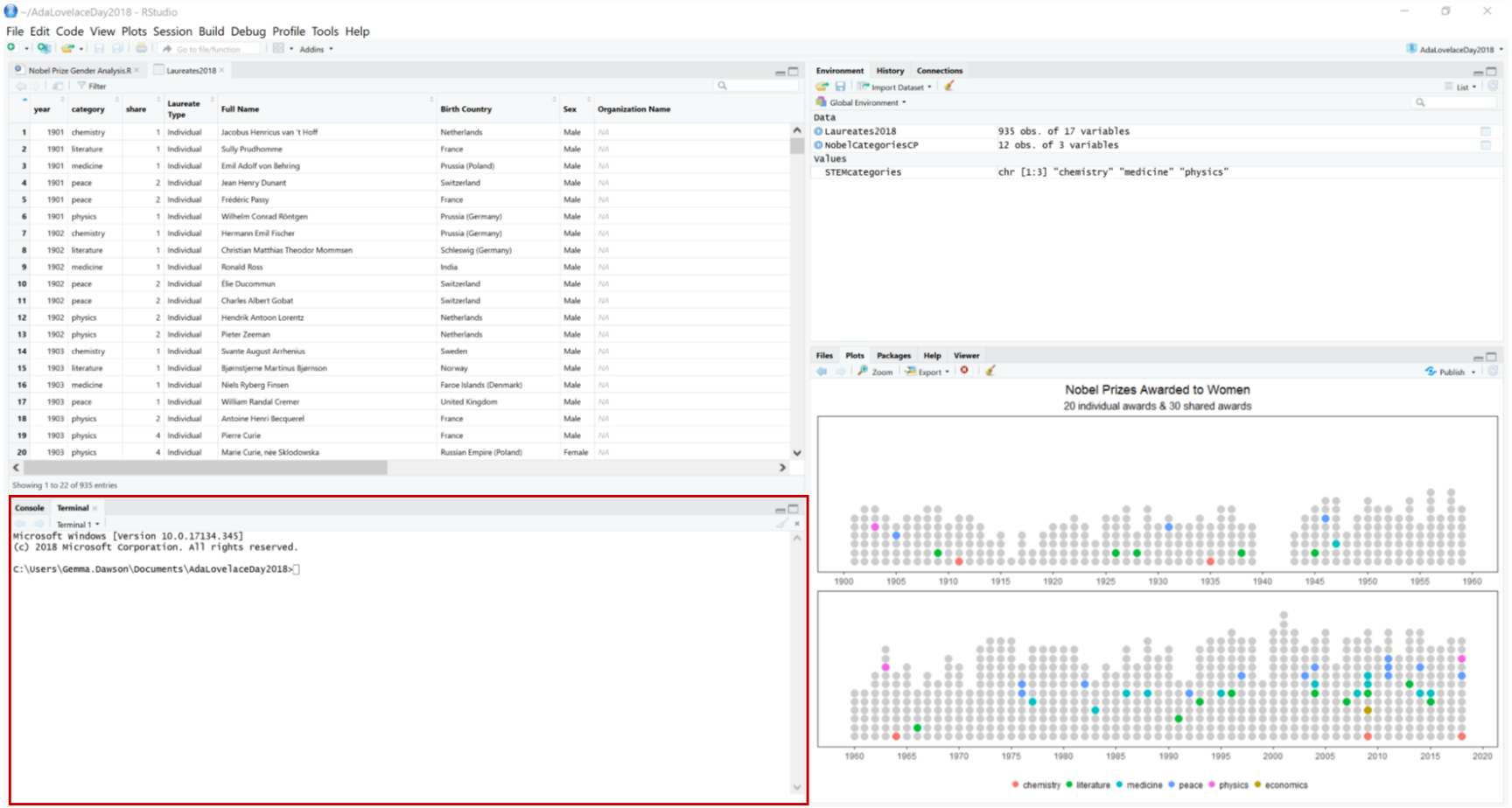
R versus Excel

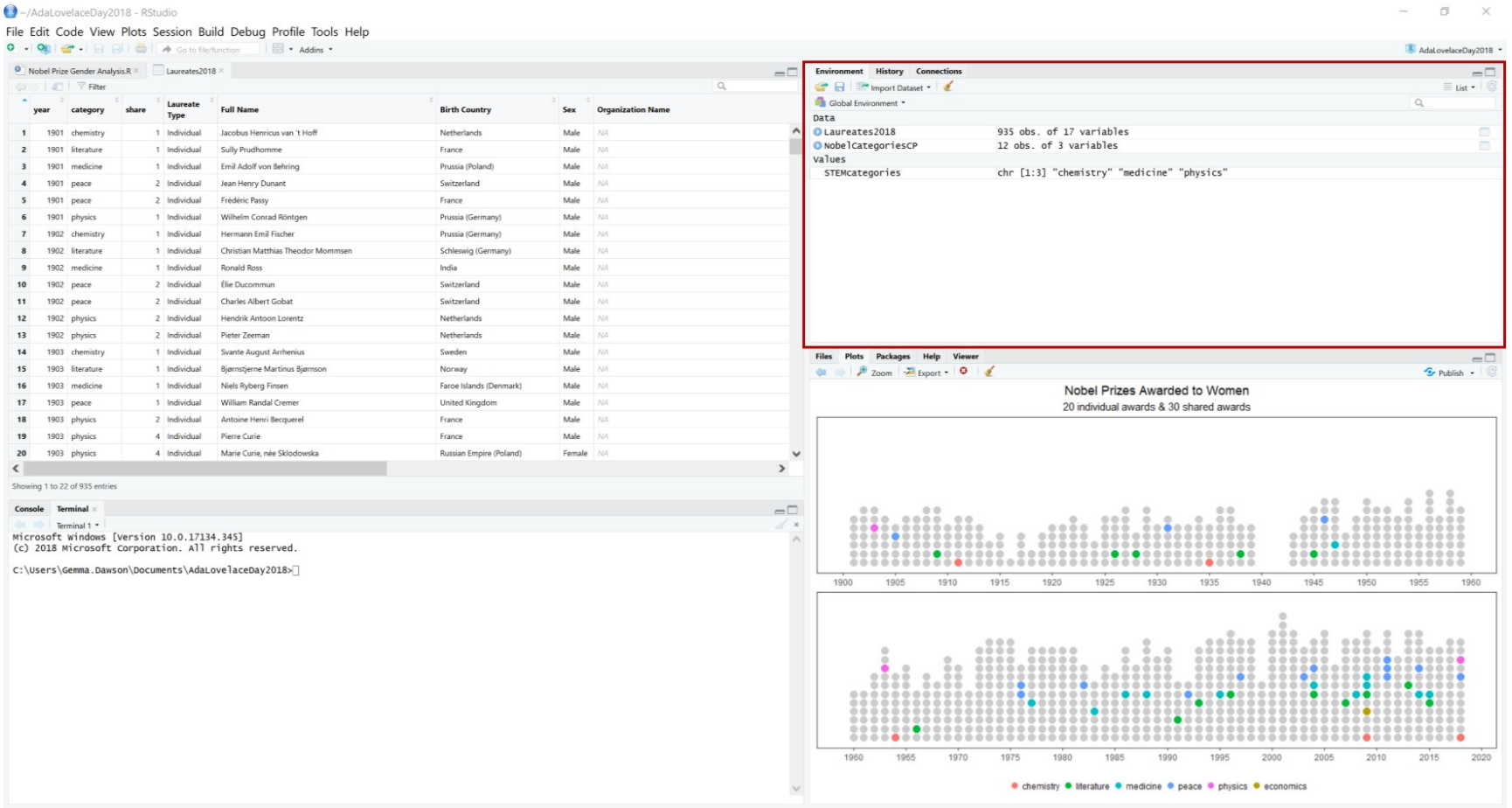

~/AdaLovelaceDay2018 - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Nobel Prize Gender Analysis.R
Laureates2018
Environment History Connections
Data
Laureates2018 935 obs. of 17 variables
NobelCategoriesCP 12 obs. of 3 variables
Values
STEMcategories chr [1:3] "chemistry" "medicine" "physics"
Console
> view(Laureates2018)
chemistry literature medicine peace physics economics

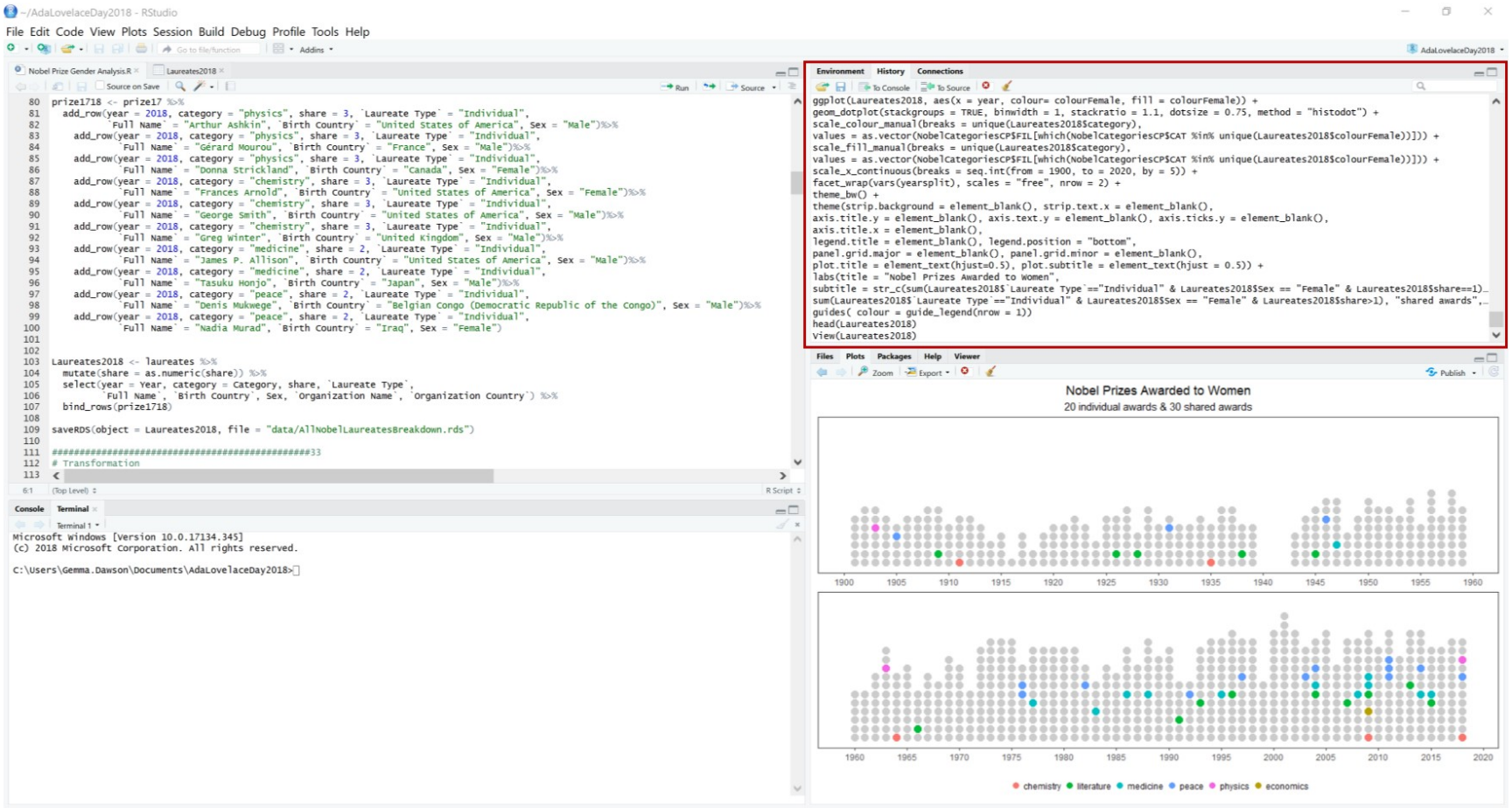

~/AdaLovelaceDay2018 - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Nobel Prize Gender Analysis.R
Laureates2018
# Nobel Prize Gender Analysis
library(pacman)
p_load(tidyverse)
prizes <- read_csv(file = "data/Nobelprizes.csv")
laureates <- read_csv(file = "data/Nobelarchive.csv")
# Grab info for 2017 from prizes
prize17 <- prizes %>% filter(year == 2017)
# tidy
laureates$Category <- tolower(laureates$Category)
laureates$share <- str_split(string = laureates$`Prize Share`, pattern = "/", 2, simplify = T)[,2]
prize17 <- prize17 %>%
mutate(`Laureate Type` = if_else(condition = is.na(surname), true = "Organization", false = "Individual"))
prize17 <- prize17 %>%
mutate(`Full Name` = if_else(condition = `Laureate Type` == "Individual", true = str_c(firstname, surname, sep = " "), false = "NA"),
`Organization Name` = if_else(condition = `Laureate Type` == "Organization", true = firstname, false = "NA"))
# Add missing columns
prize17 <- prize17 %>%
mutate(`Birth Country` = NA, `Organization Country` = NA, Sex = NA)
# Add 2017 prize winner info
# 1. Weiss
prize17$`Birth Country`[which(prize17$surname == "Weiss")] <- "Germany"
prize17$Sex[which(prize17$surname == "Weiss")] <- "Male"
Console
Terminal
wards", sep = " ")) +
guides(colour = guide_legend(nrow = 1))
> head(Laureates2018)
> view(Laureates2018)
Environment
History
Connections
Data
Laureates2018 935 obs. of 17 variables
NobelCategoriesCP 12 obs. of 3 variables
Values
STEMcategories chr [1:3] "chemistry" "medicine" "physics"
Files
Plots
Packages
Help
Viewer
Nobel Prizes Awarded to Women
20 individual awards & 30 shared awards
chemistry literature medicine peace physics economics

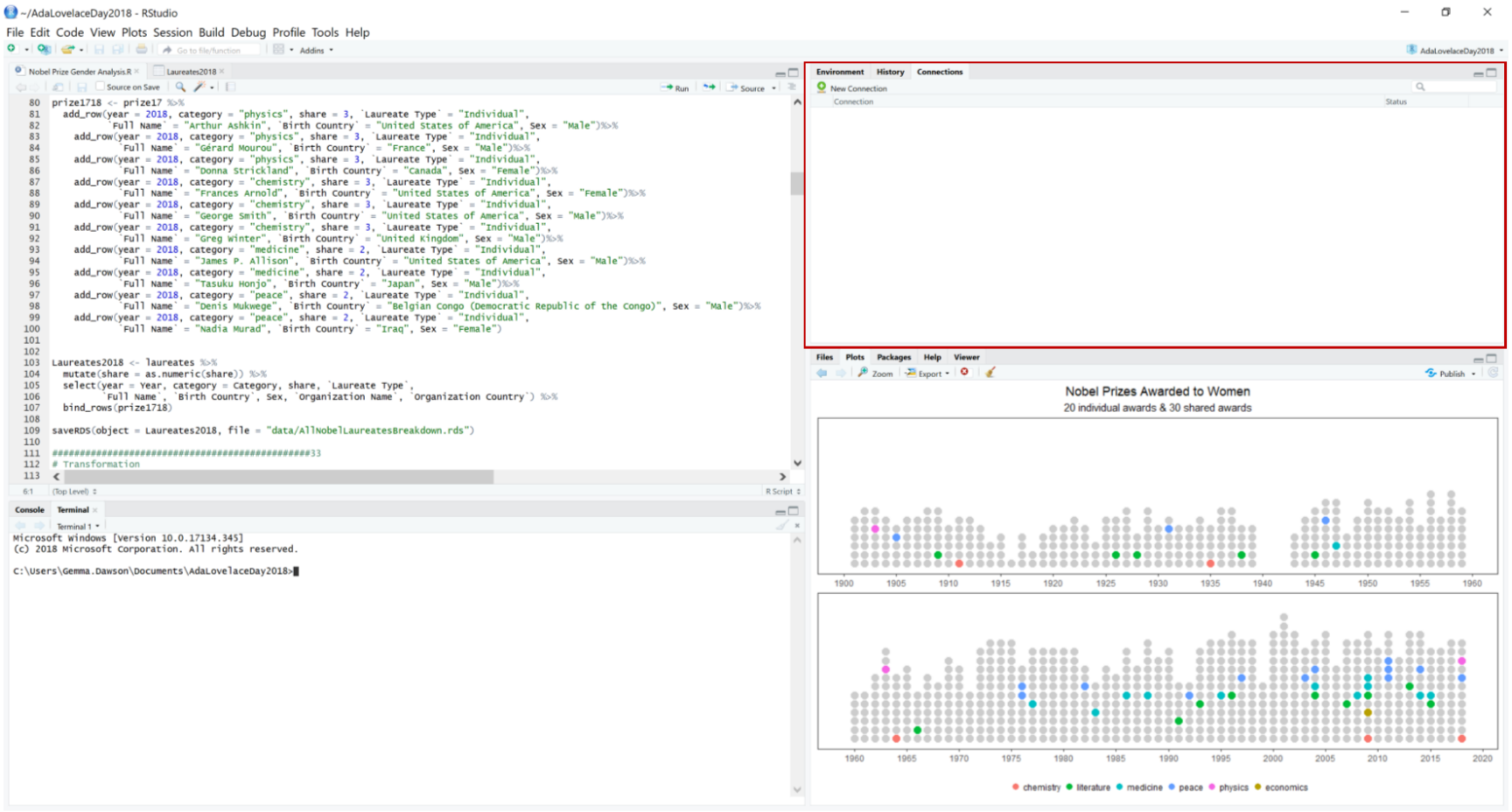

~/AdaLovelaceDay2018 - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Nobel Prize Gender Analysis.R
Laureates2018
Filter
year category share Laureate Type Full Name Birth Country Sex Organization Name
1 1901 chemistry 1 Individual Jacobus Henricus van 't Hoff Netherlands Male NA
2 1901 literature 1 Individual Sully Prudhomme France Male NA
3 1901 medicine 1 Individual Emil Adolf von Behring Prussia (Poland) Male NA
4 1901 peace 2 Individual Jean Henry Dunant Switzerland Male NA
5 1901 peace 2 Individual Frédéric Passy France Male NA
6 1901 physics 1 Individual Wilhelm Conrad Röntgen Prussia (Germany) Male NA
7 1902 chemistry 1 Individual Hermann Emil Fischer Prussia (Germany) Male NA
8 1902 literature 1 Individual Christian Matthias Theodor Mommsen Schleswig (Germany) Male NA
9 1902 medicine 1 Individual Ronald Ross India Male NA
10 1902 peace 2 Individual Élie Ducommun Switzerland Male NA
11 1902 peace 2 Individual Charles Albert Gobat Switzerland Male NA
12 1902 physics 2 Individual Hendrik Antoon Lorentz Netherlands Male NA
13 1902 physics 2 Individual Pieter Zeeman Netherlands Male NA
14 1903 chemistry 1 Individual Svante August Arrhenius Sweden Male NA
15 1903 literature 1 Individual Bjørnstjerne Martinus Bjørnson Norway Male NA
16 1903 medicine 1 Individual Niels Ryberg Finsen Faroe Islands (Denmark) Male NA
17 1903 peace 1 Individual William Randal Cremer United Kingdom Male NA
18 1903 physics 2 Individual Antoine Henri Becquerel France Male NA
19 1903 physics 4 Individual Pierre Curie France Male NA
20 1903 physics 4 Individual Marie Curie, née Sklodowska Russian Empire (Poland) Female NA
Showing 1 to 22 of 935 entries
Console Terminal
~/AdaLovelaceDay2018/
+ sum(Laureates2018$`Laureate Type`=="Individual" & Laureates2018$Sex == "Female" & Laureates2018$share>1), "shared awards", sep = " ")) +
+ guides(colour = guide_legend(nrow = 1))
> head(Laureates2018)
A tibble: 6 x 17
year category share `Laureate Type` `Full Name` `Birth Country` Sex `Organization N~ `Organization C~ yearsplit colourIndividual colourFemale colourSTEM
<dbl> <chr> <dbl> <chr> <chr> <chr> <chr> <chr> <chr> <lgl> <chr> <chr>
1 1901 chemistry 1 Individual Jacobus Henr~ Netherlands Male NA NA FALSE chemistry chemistryx chemistry
2 1901 literatu~ 1 Individual Sully Prudho~ France Male NA NA FALSE literature literaturex literatur~
3 1901 medicine 1 Individual Emil Adolf v~ Prussia (Polan~ Male NA NA FALSE medicine medicinex medicine
4 1901 peace 2 Individual Jean Henry D~ Switzerland Male NA NA FALSE peace peacex peacex
5 1901 peace 2 Individual Frédéric Pas~ France Male NA NA FALSE peace peacex peacex
6 1901 physics 1 Individual Wilhelm Conr~ Prussia (Germa~ Male NA NA FALSE physics physicsx physics
... with 4 more variables: colourFemaleSTEM <chr>, AfricanBorn <lgl>, colourAfricanBorn <chr>, colourAfricanBornSTEM <chr>
> view(Laureates2018)
Environment History Connections
Import Dataset
Global Environment
Data
Laureates2018 935 obs. of 17 variables
NobelCategoriesCP 12 obs. of 3 variables
Values
STEMcategories chr [1:3] "chemistry" "medicine" "physics"
Files Plots Packages Help Viewer
Zoom Export Publish
Nobel Prizes Awarded to Women
20 individual awards & 30 shared awards
1900 1905 1910 1915 1920 1925 1930 1935 1940 1945 1950 1955 1960
1960 1965 1970 1975 1980 1985 1990 1995 2000 2005 2010 2015 2020
chemistry literature medicine peace physics economics

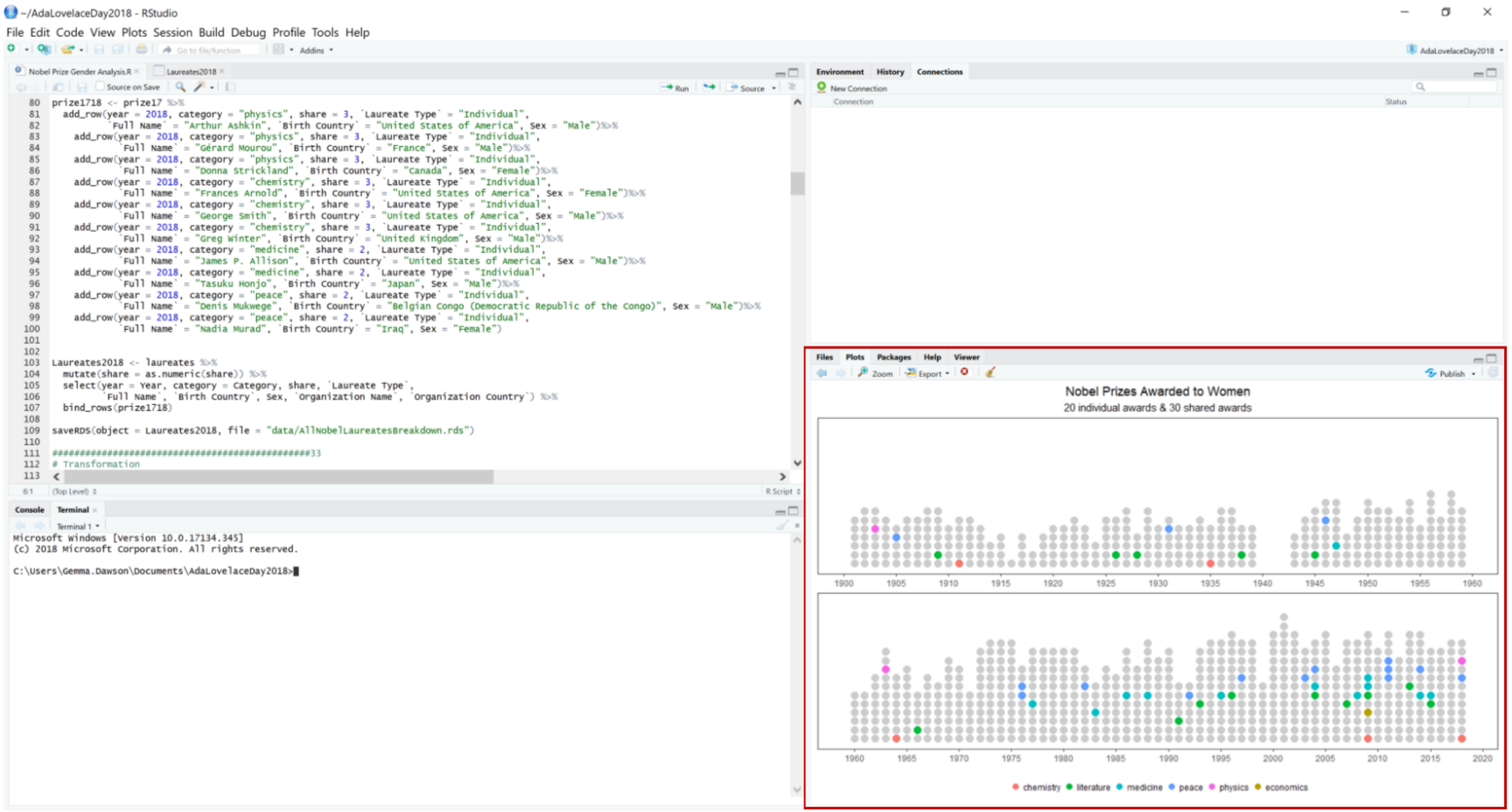

~/AdaLovelaceDay2018 - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Nobel Prize Gender Analysis.R
Laureates2018
Filter
year category share Laureate Type Full Name Birth Country Sex Organization Name
1 1901 chemistry 1 Individual Jacobus Henricus van 't Hoff Netherlands Male NA
2 1901 literature 1 Individual Sully Prudhomme France Male NA
3 1901 medicine 1 Individual Emil Adolf von Behring Prussia (Poland) Male NA
4 1901 peace 2 Individual Jean Henry Dunant Switzerland Male NA
5 1901 peace 2 Individual Frédéric Passy France Male NA
6 1901 physics 1 Individual Wilhelm Conrad Röntgen Prussia (Germany) Male NA
7 1902 chemistry 1 Individual Hermann Emil Fischer Prussia (Germany) Male NA
8 1902 literature 1 Individual Christian Matthias Theodor Mommsen Schleswig (Germany) Male NA
9 1902 medicine 1 Individual Ronald Ross India Male NA
10 1902 peace 2 Individual Élie Ducommun Switzerland Male NA
11 1902 peace 2 Individual Charles Albert Gobat Switzerland Male NA
12 1902 physics 2 Individual Hendrik Antoon Lorentz Netherlands Male NA
13 1902 physics 2 Individual Pieter Zeeman Netherlands Male NA
14 1903 chemistry 1 Individual Svante August Arrhenius Sweden Male NA
15 1903 literature 1 Individual Bjørnstjerne Martinus Bjørnson Norway Male NA
16 1903 medicine 1 Individual Niels Ryberg Finsen Faroe Islands (Denmark) Male NA
17 1903 peace 1 Individual William Randal Cremer United Kingdom Male NA
18 1903 physics 2 Individual Antoine Henri Becquerel France Male NA
19 1903 physics 4 Individual Pierre Curie France Male NA
20 1903 physics 4 Individual Marie Curie, née Sklodowska Russian Empire (Poland) Female NA
Showing 1 to 22 of 935 entries
Console Terminal
~/AdaLovelaceDay2018/
+ sum(Laureates2018$`Laureate Type`=="Individual" & Laureates2018$Sex == "Female" & Laureates2018$share>1), "shared awards", sep = " ")) +
+ guides(colour = guide_legend(nrow = 1))
> head(Laureates2018)
A tibble: 6 x 17
year category share `Laureate Type` `Full Name` `Birth Country` Sex `Organization N~ `Organization C~ yearsplit colourIndividual colourFemale colourSTEM
<dbl> <chr> <dbl> <chr> <chr> <chr> <chr> <chr> <chr> <lgl> <chr> <chr> <chr>
1 1901 chemistry 1 Individual Jacobus Henr~ Netherlands Male NA NA FALSE chemistry chemistryx chemistry
2 1901 literatu~ 1 Individual Sully Prudho~ France Male NA NA FALSE literature literaturex literatur~
3 1901 medicine 1 Individual Emil Adolf v~ Prussia (Polan~ Male NA NA FALSE medicine medicinex medicine
4 1901 peace 2 Individual Jean Henry D~ Switzerland Male NA NA FALSE peace peacex peacex
5 1901 peace 2 Individual Frédéric Pas~ France Male NA NA FALSE peace peacex peacex
6 1901 physics 1 Individual Wilhelm Conr~ Prussia (Germa~ Male NA NA FALSE physics physicsx physics
... with 4 more variables: colourFemaleSTEM <chr>, AfricanBorn <lgl>, colourAfricanBorn <chr>, colourAfricanBornSTEM <chr>
> view(Laureates2018)
Environment History Connections
Import Dataset
Global Environment
Data
Laureates2018 935 obs. of 17 variables
NobelCategoriesCP 12 obs. of 3 variables
Values
STEMcategories chr [1:3] "chemistry" "medicine" "physics"
Files Plots Packages Help Viewer
Zoom Export Publish
Nobel Prizes Awarded to Women
20 individual awards & 30 shared awards
1900 1905 1910 1915 1920 1925 1930 1935 1940 1945 1950 1955 1960
1960 1965 1970 1975 1980 1985 1990 1995 2000 2005 2010 2015 2020
chemistry literature medicine peace physics economics

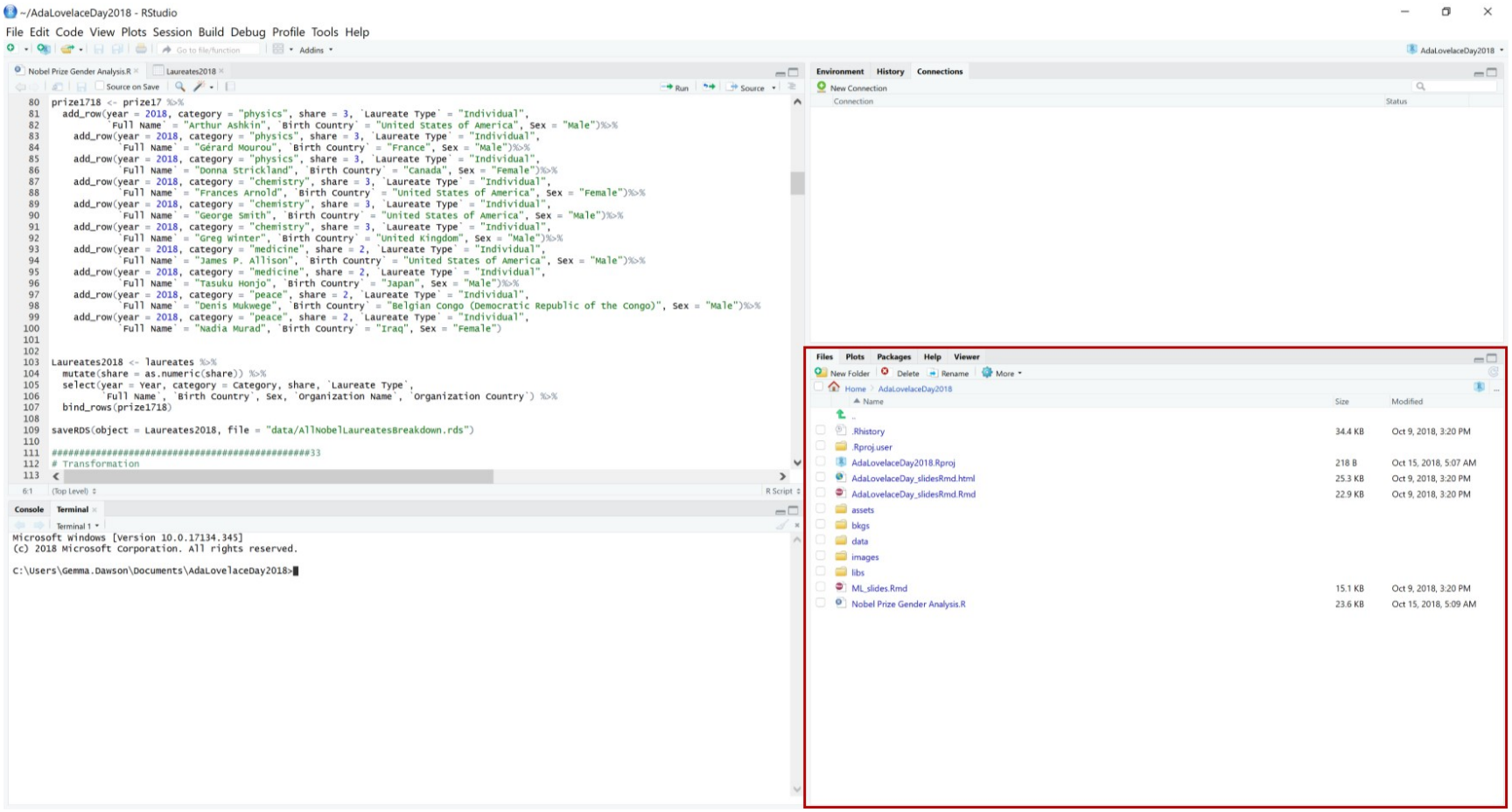

~/AdaLovelaceDay2018 - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Nobel Prize Gender Analysis.R
Laureates2018
Showing 1 to 22 of 935 entries
Console
Terminal
Microsoft Windows [Version 10.0.17134.345]
(c) 2018 Microsoft Corporation. All rights reserved.
C:\Users\Gemma.Dawson\Documents\AdaLovelaceDay2018>
Environment
History
Connections
Data
Laureates2018 935 obs. of 17 variables
NobelCategoriesCP 12 obs. of 3 variables
Values
STEMcategories chr [1:3] "chemistry" "medicine" "physics"
Files
Plots
Packages
Help
Viewer
Nobel Prizes Awarded to Women
20 individual awards & 30 shared awards
chemistry
literature
medicine
peace
physics
economics

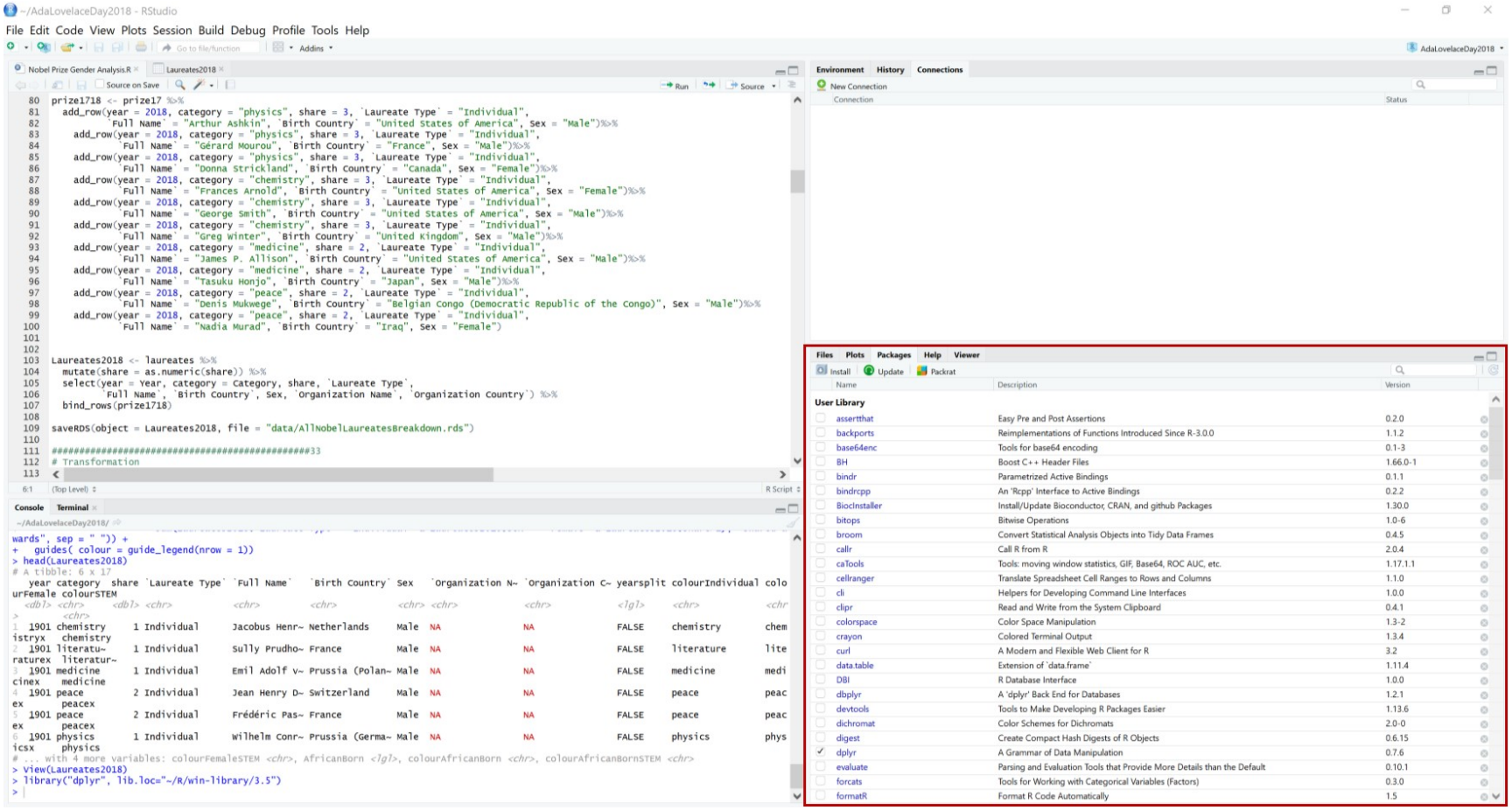

~/AdaLovelaceDay2018 - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Nobel Prize Gender Analysis.R
Laureates2018
Filter
year category share Laureate Type Full Name Birth Country Sex Organization Name
1 1901 chemistry 1 Individual Jacobus Henricus van 't Hoff Netherlands Male NA
2 1901 literature 1 Individual Sully Prudhomme France Male NA
3 1901 medicine 1 Individual Emil Adolf von Behring Prussia (Poland) Male NA
4 1901 peace 2 Individual Jean Henry Dunant Switzerland Male NA
5 1901 peace 2 Individual Frédéric Passy France Male NA
6 1901 physics 1 Individual Wilhelm Conrad Röntgen Prussia (Germany) Male NA
7 1902 chemistry 1 Individual Hermann Emil Fischer Prussia (Germany) Male NA
8 1902 literature 1 Individual Christian Matthias Theodor Mommsen Schleswig (Germany) Male NA
9 1902 medicine 1 Individual Ronald Ross India Male NA
10 1902 peace 2 Individual Élie Ducommun Switzerland Male NA
11 1902 peace 2 Individual Charles Albert Gobat Switzerland Male NA
12 1902 physics 2 Individual Hendrik Antoon Lorentz Netherlands Male NA
13 1902 physics 2 Individual Pieter Zeeman Netherlands Male NA
14 1903 chemistry 1 Individual Svante August Arrhenius Sweden Male NA
15 1903 literature 1 Individual Bjørnstjerne Martinus Bjørnson Norway Male NA
16 1903 medicine 1 Individual Niels Ryberg Finsen Faroe Islands (Denmark) Male NA
17 1903 peace 1 Individual William Randal Cremer United Kingdom Male NA
18 1903 physics 2 Individual Antoine Henri Becquerel France Male NA
19 1903 physics 4 Individual Pierre Curie France Male NA
20 1903 physics 4 Individual Marie Curie, née Sklodowska Russian Empire (Poland) Female NA
Showing 1 to 22 of 935 entries
Console Terminal
Microsoft Windows [Version 10.0.17134.345]
(c) 2018 Microsoft Corporation. All rights reserved.
C:\Users\Gemma.Dawson\Documents\AdaLovelaceDay2018>
Environment History Connections
Import Dataset
Global Environment
Data
Laureates2018 935 obs. of 17 variables
NobelCategoriesCP 12 obs. of 3 variables
Values
STEMcategories chr [1:3] "chemistry" "medicine" "physics"
Files Plots Packages Help Viewer
Zoom Export Publish
Nobel Prizes Awarded to Women
20 individual awards & 30 shared awards
1900 1905 1910 1915 1920 1925 1930 1935 1940 1945 1950 1955 1960
1960 1965 1970 1975 1980 1985 1990 1995 2000 2005 2010 2015 2020
chemistry literature medicine peace physics economics

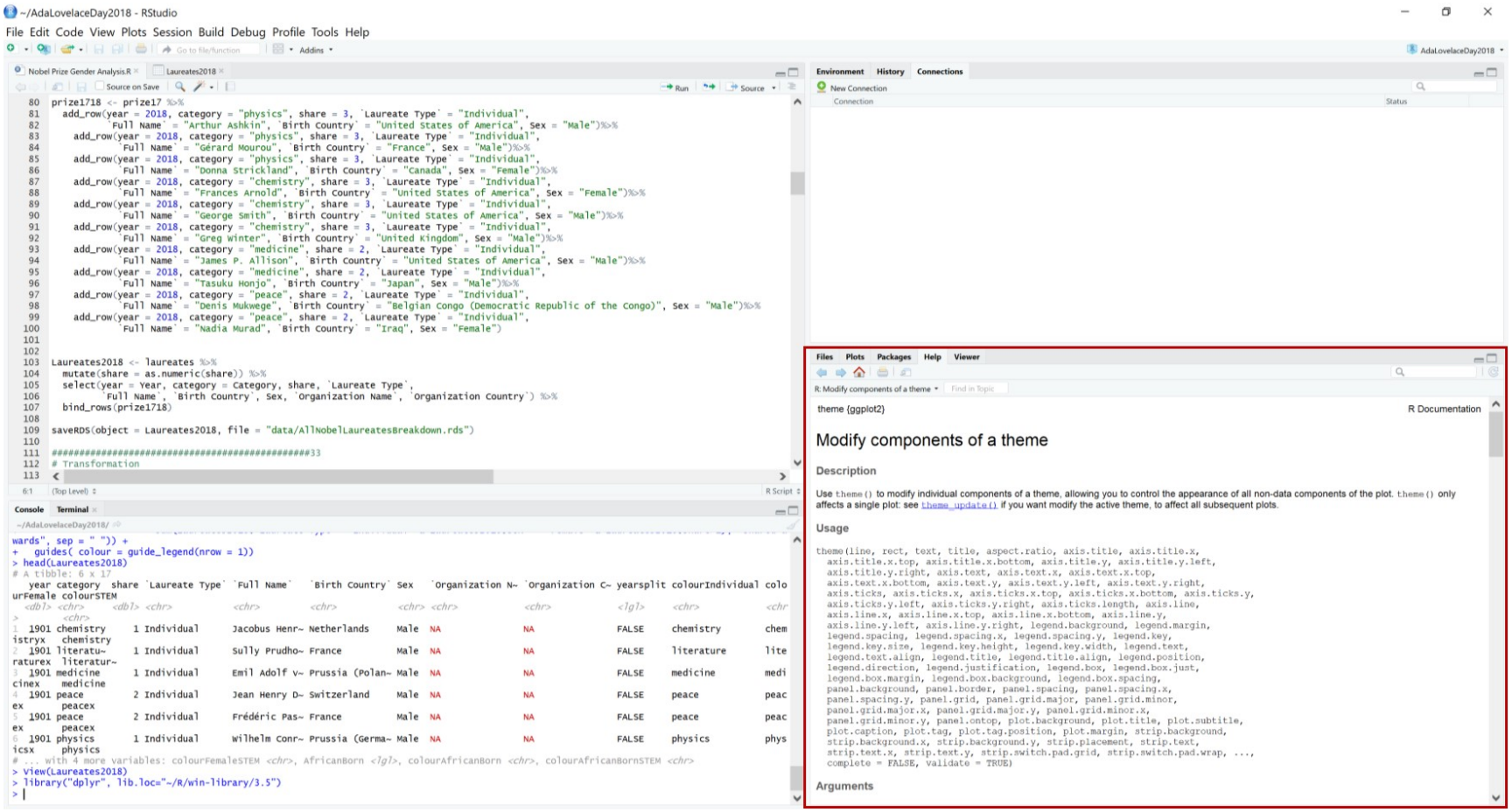

~/AdaLovelaceDay2018 - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Nobel Prize Gender Analysis.R
Laureates2018
prize1718 <- prize17 %>%
add_row(year = 2018, category = "physics", share = 3, `Laureate Type` = "Individual",
`Full Name` = "Arthur Ashkin", `Birth Country` = "United States of America", Sex = "Male")%>%
add_row(year = 2018, category = "physics", share = 3, `Laureate Type` = "Individual",
`Full Name` = "Gérard Mourou", `Birth Country` = "France", Sex = "Male")%>%
add_row(year = 2018, category = "physics", share = 3, `Laureate Type` = "Individual",
`Full Name` = "Donna Strickland", `Birth Country` = "Canada", Sex = "Female")%>%
add_row(year = 2018, category = "chemistry", share = 3, `Laureate Type` = "Individual",
`Full Name` = "Frances Arnold", `Birth Country` = "United States of America", Sex = "Female")%>%
add_row(year = 2018, category = "chemistry", share = 3, `Laureate Type` = "Individual",
`Full Name` = "George Smith", `Birth Country` = "United States of America", Sex = "Male")%>%
add_row(year = 2018, category = "chemistry", share = 3, `Laureate Type` = "Individual",
`Full Name` = "Greg Winter", `Birth Country` = "United Kingdom", Sex = "Male")%>%
add_row(year = 2018, category = "medicine", share = 2, `Laureate Type` = "Individual",
`Full Name` = "James P. Allison", `Birth Country` = "United States of America", Sex = "Male")%>%
add_row(year = 2018, category = "medicine", share = 2, `Laureate Type` = "Individual",
`Full Name` = "Tasuku Honjo", `Birth Country` = "Japan", Sex = "Male")%>%
add_row(year = 2018, category = "peace", share = 2, `Laureate Type` = "Individual",
`Full Name` = "Denis Mukwege", `Birth Country` = "Belgian Congo (Democratic Republic of the Congo)", Sex = "Male")%>%
add_row(year = 2018, category = "peace", share = 2, `Laureate Type` = "Individual",
`Full Name` = "Nadia Murad", `Birth Country` = "Iraq", Sex = "Female")
Laureates2018 <- laureates %>%
mutate(share = as.numeric(share)) %>%
select(year = Year, category = Category, share, `Laureate Type`,
`Full Name`, `Birth Country`, Sex, `Organization Name`, `Organization Country`) %>%
bind_rows(prize1718)
saveRDS(object = Laureates2018, file = "data/AllNobelLaureatesBreakdown.rds")
##33
Transformation
Console
Terminal
Microsoft Windows [Version 10.0.17134.345]
(c) 2018 Microsoft Corporation. All rights reserved.
C:\Users\Gemma.Dawson\Documents\AdaLovelaceDay2018>
Environment
History
Connections
ggplot(Laureates2018, aes(x = year, colour= colourFemale, fill = colourFemale)) +
geom_dotplot(stackgroups = TRUE, binwidth = 1, stackratio = 1.1, dotsize = 0.75, method = "histodot") +
scale_colour_manual(breaks = unique(Laureates2018$category),
values = as.vector(NobelCategoriesCP$FIL[which(NobelCategoriesCP$CAT %in% unique(Laureates2018$colourFemale))])) +
scale_fill_manual(breaks = unique(Laureates2018$category),
values = as.vector(NobelCategoriesCP$FIL[which(NobelCategoriesCP$CAT %in% unique(Laureates2018$colourFemale))])) +
scale_x_continuous(breaks = seq.int(from = 1900, to = 2020, by = 5)) +
facet_wrap(vars(yearsplit), scales = "free", nrow = 2) +
theme_bw() +
theme(strip.background = element_blank(), strip.text.x = element_blank(),
axis.title.y = element_blank(), axis.text.y = element_blank(), axis.ticks.y = element_blank(),
axis.title.x = element_blank(),
legend.title = element_blank(), legend.position = "bottom",
panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
plot.title = element_text(hjust=0.5), plot.subtitle = element_text(hjust = 0.5)) +
labs(title = "Nobel Prizes Awarded to Women",
guides(colour = guide_legend(nrow = 1))
head(Laureates2018)
View(Laureates2018)
Files
Plots
Packages
Help
Viewer
Nobel Prizes Awarded to Women
20 individual awards & 30 shared awards
chemistry
literature
medicine
peace
physics
economics


~/AdaLovelaceDay2018 - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Nobel Prize Gender Analysis.R
Laureates2018
prize1718 <- prize17 %>%
add_row(year = 2018, category = "physics", share = 3, `Laureate Type` = "Individual",
`Full Name` = "Arthur Ashkin", `Birth Country` = "United States of America", Sex = "Male")%>%
add_row(year = 2018, category = "physics", share = 3, `Laureate Type` = "Individual",
`Full Name` = "Gérard Mourou", `Birth Country` = "France", Sex = "Male")%>%
add_row(year = 2018, category = "physics", share = 3, `Laureate Type` = "Individual",
`Full Name` = "Donna Strickland", `Birth Country` = "Canada", Sex = "Female")%>%
add_row(year = 2018, category = "chemistry", share = 3, `Laureate Type` = "Individual",
`Full Name` = "Frances Arnold", `Birth Country` = "United States of America", Sex = "Female")%>%
add_row(year = 2018, category = "chemistry", share = 3, `Laureate Type` = "Individual",
`Full Name` = "George Smith", `Birth Country` = "United States of America", Sex = "Male")%>%
add_row(year = 2018, category = "chemistry", share = 3, `Laureate Type` = "Individual",
`Full Name` = "Greg Winter", `Birth Country` = "United Kingdom", Sex = "Male")%>%
add_row(year = 2018, category = "medicine", share = 2, `Laureate Type` = "Individual",
`Full Name` = "James P. Allison", `Birth Country` = "United States of America", Sex = "Male")%>%
add_row(year = 2018, category = "medicine", share = 2, `Laureate Type` = "Individual",
`Full Name` = "Tasuku Honjo", `Birth Country` = "Japan", Sex = "Male")%>%
add_row(year = 2018, category = "peace", share = 2, `Laureate Type` = "Individual",
`Full Name` = "Denis Mukwege", `Birth Country` = "Belgian Congo (Democratic Republic of the Congo)", Sex = "Male")%>%
add_row(year = 2018, category = "peace", share = 2, `Laureate Type` = "Individual",
`Full Name` = "Nadia Murad", `Birth Country` = "Iraq", Sex = "Female")
Laureates2018 <- laureates %>%
mutate(share = as.numeric(share)) %>%
select(year = Year, category = Category, share, `Laureate Type`,
`Full Name`, `Birth Country`, Sex, `Organization Name`, `Organization Country`) %>%
bind_rows(prize1718)
saveRDS(object = Laureates2018, file = "data/AllNobelLaureatesBreakdown.rds")
##33
Transformation
Console
Terminal
Microsoft Windows [Version 10.0.17134.345]
(c) 2018 Microsoft Corporation. All rights reserved.
C:\Users\Gemma.Dawson\Documents\AdaLovelaceDay2018>
Environment
History
Connections
New Connection
Connection
Status
Files
Plots
Packages
Help
Viewer
Nobel Prizes Awarded to Women
20 individual awards & 30 shared awards
1900 1905 1910 1915 1920 1925 1930 1935 1940 1945 1950 1955 1960
1960 1965 1970 1975 1980 1985 1990 1995 2000 2005 2010 2015 2020
chemistry literature medicine peace physics economics


~/AdaLovelaceDay2018 - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Nobel Prize Gender Analysis.R
Laureates2018
prize1718 <- prize17 %>%
add_row(year = 2018, category = "physics", share = 3, `Laureate Type` = "Individual",
`Full Name` = "Arthur Ashkin", `Birth Country` = "United States of America", Sex = "Male")%>%
add_row(year = 2018, category = "physics", share = 3, `Laureate Type` = "Individual",
`Full Name` = "Gérard Mourou", `Birth Country` = "France", Sex = "Male")%>%
add_row(year = 2018, category = "physics", share = 3, `Laureate Type` = "Individual",
`Full Name` = "Donna Strickland", `Birth Country` = "Canada", Sex = "Female")%>%
add_row(year = 2018, category = "chemistry", share = 3, `Laureate Type` = "Individual",
`Full Name` = "Frances Arnold", `Birth Country` = "United States of America", Sex = "Female")%>%
add_row(year = 2018, category = "chemistry", share = 3, `Laureate Type` = "Individual",
`Full Name` = "George Smith", `Birth Country` = "United States of America", Sex = "Male")%>%
add_row(year = 2018, category = "chemistry", share = 3, `Laureate Type` = "Individual",
`Full Name` = "Greg Winter", `Birth Country` = "United Kingdom", Sex = "Male")%>%
add_row(year = 2018, category = "medicine", share = 2, `Laureate Type` = "Individual",
`Full Name` = "James P. Allison", `Birth Country` = "United States of America", Sex = "Male")%>%
add_row(year = 2018, category = "medicine", share = 2, `Laureate Type` = "Individual",
`Full Name` = "Tasuku Honjo", `Birth Country` = "Japan", Sex = "Male")%>%
add_row(year = 2018, category = "peace", share = 2, `Laureate Type` = "Individual",
`Full Name` = "Denis Mukwege", `Birth Country` = "Belgian Congo (Democratic Republic of the Congo)", Sex = "Male")%>%
add_row(year = 2018, category = "peace", share = 2, `Laureate Type` = "Individual",
`Full Name` = "Nadia Murad", `Birth Country` = "Iraq", Sex = "Female")
Laureates2018 <- laureates %>%
mutate(share = as.numeric(share)) %>%
select(year = Year, category = Category, share, `Laureate Type`,
`Full Name`, `Birth Country`, Sex, `Organization Name`, `Organization Country`) %>%
bind_rows(prize1718)
saveRDS(object = Laureates2018, file = "data/AllNobelLaureatesBreakdown.rds")
##33
Transformation
Environment History Connections
New Connection
Connection
Status
Console Terminal
Microsoft Windows [Version 10.0.17134.345]
(c) 2018 Microsoft Corporation. All rights reserved.
C:\Users\Gemma.Dawson\Documents\AdaLovelaceDay2018>
Files Plots Packages Help Viewer
Nobel Prizes Awarded to Women
20 individual awards & 30 shared awards
1900 1905 1910 1915 1920 1925 1930 1935 1940 1945 1950 1955 1960
1960 1965 1970 1975 1980 1985 1990 1995 2000 2005 2010 2015 2020
chemistry literature medicine peace physics economics


~/AdaLovelaceDay2018 - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Nobel Prize Gender Analysis.R
Laureates2018
Source on Save
Run
Source
prize1718 <- prize17 %>%
add_row(year = 2018, category = "physics", share = 3, `Laureate Type` = "Individual",
`Full Name` = "Arthur Ashkin", `Birth Country` = "United States of America", Sex = "Male")%>%
add_row(year = 2018, category = "physics", share = 3, `Laureate Type` = "Individual",
`Full Name` = "Gérard Mourou", `Birth Country` = "France", Sex = "Male")%>%
add_row(year = 2018, category = "physics", share = 3, `Laureate Type` = "Individual",
`Full Name` = "Donna Strickland", `Birth Country` = "Canada", Sex = "Female")%>%
add_row(year = 2018, category = "chemistry", share = 3, `Laureate Type` = "Individual",
`Full Name` = "Frances Arnold", `Birth Country` = "United States of America", Sex = "Female")%>%
add_row(year = 2018, category = "chemistry", share = 3, `Laureate Type` = "Individual",
`Full Name` = "George Smith", `Birth Country` = "United States of America", Sex = "Male")%>%
add_row(year = 2018, category = "chemistry", share = 3, `Laureate Type` = "Individual",
`Full Name` = "Greg Winter", `Birth Country` = "United Kingdom", Sex = "Male")%>%
add_row(year = 2018, category = "medicine", share = 2, `Laureate Type` = "Individual",
`Full Name` = "James P. Allison", `Birth Country` = "United States of America", Sex = "Male")%>%
add_row(year = 2018, category = "medicine", share = 2, `Laureate Type` = "Individual",
`Full Name` = "Tasuku Honjo", `Birth Country` = "Japan", Sex = "Male")%>%
add_row(year = 2018, category = "peace", share = 2, `Laureate Type` = "Individual",
`Full Name` = "Denis Mukwege", `Birth Country` = "Belgian Congo (Democratic Republic of the Congo)", Sex = "Male")%>%
add_row(year = 2018, category = "peace", share = 2, `Laureate Type` = "Individual",
`Full Name` = "Nadia Murad", `Birth Country` = "Iraq", Sex = "Female")
Laureates2018 <- laureates %>%
mutate(share = as.numeric(share)) %>%
select(year = Year, category = Category, share, `Laureate Type`,
`Full Name`, `Birth Country`, Sex, `Organization Name`, `Organization Country`) %>%
bind_rows(prize1718)
saveRDS(object = Laureates2018, file = "data/AllNobelLaureatesBreakdown.rds")
##33
Transformation
6:1
(Top Level)
R Script
Console
Terminal
Terminal 1
Microsoft Windows [Version 10.0.17134.345]
(c) 2018 Microsoft Corporation. All rights reserved.
C:\Users\Gemma.Dawson\Documents\AdaLovelaceDay2018>
Environment
History
Connections
New Connection
Connection
Status
AdaLovelaceDay2018
Files
Plots
Packages
Help
Viewer
New Folder
Delete
Rename
More
Home
AdaLovelaceDay2018
Name
Size
Modified
.Rhistory 34.4 KB Oct 9, 2018, 3:20 PM
.Rproj.user
AdaLovelaceDay2018.Rproj 218 B Oct 15, 2018, 5:07 AM
AdaLovelaceDay_slidesRmd.html 25.3 KB Oct 9, 2018, 3:20 PM
AdaLovelaceDay_slidesRmd.Rmd 22.9 KB Oct 9, 2018, 3:20 PM
assets
bkgs
data
images
libs
ML_slides.Rmd 15.1 KB Oct 9, 2018, 3:20 PM
Nobel Prize Gender Analysis.R 23.6 KB Oct 15, 2018, 5:09 AM


~/AdaLovelaceDay2018 - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Nobel Prize Gender Analysis.R
Laureates2018
Source on Save
Run
Source
prize1718 <- prize17 %>%
add_row(year = 2018, category = "physics", share = 3, `Laureate Type` = "Individual",
`Full Name` = "Arthur Ashkin", `Birth Country` = "United States of America", Sex = "Male")%>%
add_row(year = 2018, category = "physics", share = 3, `Laureate Type` = "Individual",
`Full Name` = "Gérard Mourou", `Birth Country` = "France", Sex = "Male")%>%
add_row(year = 2018, category = "physics", share = 3, `Laureate Type` = "Individual",
`Full Name` = "Donna Strickland", `Birth Country` = "Canada", Sex = "Female")%>%
add_row(year = 2018, category = "chemistry", share = 3, `Laureate Type` = "Individual",
`Full Name` = "Frances Arnold", `Birth Country` = "United States of America", Sex = "Female")%>%
add_row(year = 2018, category = "chemistry", share = 3, `Laureate Type` = "Individual",
`Full Name` = "George Smith", `Birth Country` = "United States of America", Sex = "Male")%>%
add_row(year = 2018, category = "chemistry", share = 3, `Laureate Type` = "Individual",
`Full Name` = "Greg Winter", `Birth Country` = "United Kingdom", Sex = "Male")%>%
add_row(year = 2018, category = "medicine", share = 2, `Laureate Type` = "Individual",
`Full Name` = "James P. Allison", `Birth Country` = "United States of America", Sex = "Male")%>%
add_row(year = 2018, category = "medicine", share = 2, `Laureate Type` = "Individual",
`Full Name` = "Tasuku Honjo", `Birth Country` = "Japan", Sex = "Male")%>%
add_row(year = 2018, category = "peace", share = 2, `Laureate Type` = "Individual",
`Full Name` = "Denis Mukwege", `Birth Country` = "Belgian Congo (Democratic Republic of the Congo)", Sex = "Male")%>%
add_row(year = 2018, category = "peace", share = 2, `Laureate Type` = "Individual",
`Full Name` = "Nadia Murad", `Birth Country` = "Iraq", Sex = "Female")
Laureates2018 <- laureates %>%
mutate(share = as.numeric(share)) %>%
select(year = Year, category = Category, share, `Laureate Type`,
`Full Name`, `Birth Country`, Sex, `Organization Name`, `Organization Country`) %>%
bind_rows(prize1718)
saveRDS(object = Laureates2018, file = "data/AllNobelLaureatesBreakdown.rds")
##33
Transformation
6:1
(Top Level)
R Script
Console
Terminal
~/AdaLovelaceDay2018/
wards", sep = " ")) +
+ guides(colour = guide_legend(nrow = 1))
> head(Laureates2018)
A tibble: 6 x 17
year category share `Laureate Type` `Full Name` `Birth Country` Sex `Organization N~ `Organization C~ yearsplit colourIndividual colourFemale colourSTEM
<dbl> <chr> <dbl> <chr> <chr> <chr> <chr> <chr> <chr> <lgl> <chr> <chr> <chr>
1 1901 chemistry 1 Individual Jacobus Henr~ Netherlands Male NA NA FALSE chemistry chemistryx chemistry
2 1901 literatu~ 1 Individual Sully Prudho~ France Male NA NA FALSE literature literaturex literatur~
3 1901 medicine 1 Individual Emil Adolf v~ Prussia (Polan~ Male NA NA FALSE medicine medicinex medicine
4 1901 peace 2 Individual Jean Henry D~ Switzerland Male NA NA FALSE peace peacex peacex
5 1901 peace 2 Individual Frédéric Pas~ France Male NA NA FALSE peace peacex peacex
6 1901 physics 1 Individual Wilhelm Conr~ Prussia (Germa~ Male NA NA FALSE physics physicsx physics
... with 4 more variables: colourFemaleSTEM <chr>, AfricanBorn <lgl>, colourAfricanBorn <chr>, colourAfricanBornSTEM <chr>
> view(Laureates2018)
> library("dplyr", lib.loc="~/R/win-library/3.5")
>
Environment
History
Connections
New Connection
Connection
Status
Files
Plots
Packages
Help
Viewer
Install
Update
Packrat
Name
Description
Version
User Library
assertthat Easy Pre and Post Assertions 0.2.0
backports Reimplementations of Functions Introduced Since R-3.0.0 1.1.2
base64enc Tools for base64 encoding 0.1-3
BH Boost C++ Header Files 1.66.0-1
bindr Parametrized Active Bindings 0.1.1
bindrcpp An 'Rcpp' Interface to Active Bindings 0.2.2
BiocInstaller Install/Update Bioconductor, CRAN, and github Packages 1.30.0
bitops Bitwise Operations 1.0-6
broom Convert Statistical Analysis Objects into Tidy Data Frames 0.4.5
callr Call R from R 2.0.4
caTools Tools: moving window statistics, GIF, Base64, ROC AUC, etc. 1.17.1.1
cellranger Translate Spreadsheet Cell Ranges to Rows and Columns 1.1.0
cli Helpers for Developing Command Line Interfaces 1.0.0
clipr Read and Write from the System Clipboard 0.4.1
colorspace Color Space Manipulation 1.3-2
crayon Colored Terminal Output 1.3.4
curl A Modern and Flexible Web Client for R 3.2
data.table Extension of `data.frame` 1.11.4
DBI R Database Interface 1.0.0
dbplyr A 'dplyr' Back End for Databases 1.2.1
devtools Tools to Make Developing R Packages Easier 1.13.6
dichromat Color Schemes for Dichromats 2.0-0
digest Create Compact Hash Digests of R Objects 0.6.15
dplyr A Grammar of Data Manipulation 0.7.6
evaluate Parsing and Evaluation Tools that Provide More Details than the Default 0.10.1
forcats Tools for Working with Categorical Variables (Factors) 0.3.0
formatR Format R Code Automatically 1.5


~/AdaLovelaceDay2018 - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Nobel Prize Gender Analysis.R
Laureates2018
prize1718 <- prize17 %>%
add_row(year = 2018, category = "physics", share = 3, `Laureate Type` = "Individual",
`Full Name` = "Arthur Ashkin", `Birth Country` = "United States of America", Sex = "Male")%>%
add_row(year = 2018, category = "physics", share = 3, `Laureate Type` = "Individual",
`Full Name` = "Gérard Mourou", `Birth Country` = "France", Sex = "Male")%>%
add_row(year = 2018, category = "physics", share = 3, `Laureate Type` = "Individual",
`Full Name` = "Donna Strickland", `Birth Country` = "Canada", Sex = "Female")%>%
add_row(year = 2018, category = "chemistry", share = 3, `Laureate Type` = "Individual",
`Full Name` = "Frances Arnold", `Birth Country` = "United States of America", Sex = "Female")%>%
add_row(year = 2018, category = "chemistry", share = 3, `Laureate Type` = "Individual",
`Full Name` = "George Smith", `Birth Country` = "United States of America", Sex = "Male")%>%
add_row(year = 2018, category = "chemistry", share = 3, `Laureate Type` = "Individual",
`Full Name` = "Greg Winter", `Birth Country` = "United Kingdom", Sex = "Male")%>%
add_row(year = 2018, category = "medicine", share = 2, `Laureate Type` = "Individual",
`Full Name` = "James P. Allison", `Birth Country` = "United States of America", Sex = "Male")%>%
add_row(year = 2018, category = "medicine", share = 2, `Laureate Type` = "Individual",
`Full Name` = "Tasuku Honjo", `Birth Country` = "Japan", Sex = "Male")%>%
add_row(year = 2018, category = "peace", share = 2, `Laureate Type` = "Individual",
`Full Name` = "Denis Mukwege", `Birth Country` = "Belgian Congo (Democratic Republic of the Congo)", Sex = "Male")%>%
add_row(year = 2018, category = "peace", share = 2, `Laureate Type` = "Individual",
`Full Name` = "Nadia Murad", `Birth Country` = "Iraq", Sex = "Female")
Laureates2018 <- laureates %>%
mutate(share = as.numeric(share)) %>%
select(year = Year, category = Category, share, `Laureate Type`,
`Full Name`, `Birth Country`, Sex, `Organization Name`, `Organization Country`) %>%
bind_rows(prize1718)
saveRDS(object = Laureates2018, file = "data/AllNobelLaureatesBreakdown.rds")
##33
Transformation
Console
Terminal
~/AdaLovelaceDay2018/
wards", sep = " ")) +
+ guides(colour = guide_legend(nrow = 1))
> head(Laureates2018)
A tibble: 6 x 17
year category share `Laureate Type` `Full Name` `Birth Country` Sex `Organization N~ `Organization C~ yearsplit colourIndividual colourFemale colourSTEM
<dbl> <chr> <dbl> <chr> <chr> <chr> <chr> <chr> <chr> <lgl> <chr> <chr> <chr>
1 1901 chemistry 1 Individual Jacobus Henr~ Netherlands Male NA NA FALSE chemistry chemistryx chemistry
2 1901 literatu~ 1 Individual Sully Prudho~ France Male NA NA FALSE literature literaturex literatur~
3 1901 medicine 1 Individual Emil Adolf v~ Prussia (Polan~ Male NA NA FALSE medicine medicinex medicine
4 1901 peace 2 Individual Jean Henry D~ Switzerland Male NA NA FALSE peace peacex peacex
5 1901 peace 2 Individual Frédéric Pas~ France Male NA NA FALSE peace peacex peacex
6 1901 physics 1 Individual Wilhelm Conr~ Prussia (Germa~ Male NA NA FALSE physics physicsx physics
... with 4 more variables: colourFemaleSTEM <chr>, AfricanBorn <lgl>, colourAfricanBorn <chr>, colourAfricanBornSTEM <chr>
> view(Laureates2018)
> library("dplyr", lib.loc="~/R/win-library/3.5")
Environment
History
Connections
New Connection
Connection
Status
AdaLovelaceDay2018
Files
Plots
Packages
Help
Viewer
R: Modify components of a theme
Find in Topic
theme {ggplot2}
R Documentation
Modify components of a theme
Description
Use theme() to modify individual components of a theme, allowing you to control the appearance of all non-data components of the plot. theme() only affects a single plot: see theme_update() if you want modify the active theme, to affect all subsequent plots.
Usage
theme(line, rect, text, title, aspect.ratio, axis.title, axis.title.x,
axis.title.x.top, axis.title.x.bottom, axis.title.y, axis.title.y.left,
axis.title.y.right, axis.text, axis.text.x, axis.text.x.top,
axis.text.x.bottom, axis.text.y, axis.text.y.left, axis.text.y.right,
axis.ticks, axis.ticks.x, axis.ticks.x.top, axis.ticks.x.bottom, axis.ticks.y,
axis.ticks.y.left, axis.ticks.y.right, axis.ticks.length, axis.line,
axis.line.x, axis.line.x.top, axis.line.x.bottom, axis.line.y,
axis.line.y.left, axis.line.y.right, legend.background, legend.margin,
legend.spacing, legend.spacing.x, legend.spacing.y, legend.key,
legend.key.size, legend.key.height, legend.key.width, legend.text,
legend.text.align, legend.title, legend.title.align, legend.position,
legend.direction, legend.justification, legend.box, legend.box.just,
legend.box.margin, legend.box.background, legend.box.spacing,
panel.background, panel.border, panel.spacing, panel.spacing.x,
panel.spacing.y, panel.grid, panel.grid.major, panel.grid.minor,
panel.grid.major.x, panel.grid.major.y, panel.grid.minor.x,
panel.grid.minor.y, panel.ontop, plot.background, plot.title, plot.subtitle,
plot.caption, plot.tag, plot.tag.position, plot.margin, strip.background,
strip.background.x, strip.background.y, strip.placement, strip.text,
strip.text.x, strip.text.y, strip.switch.pad.grid, strip.switch.pad.wrap, ...,
complete = FALSE, validate = TRUE)
Arguments

The Community

The Community

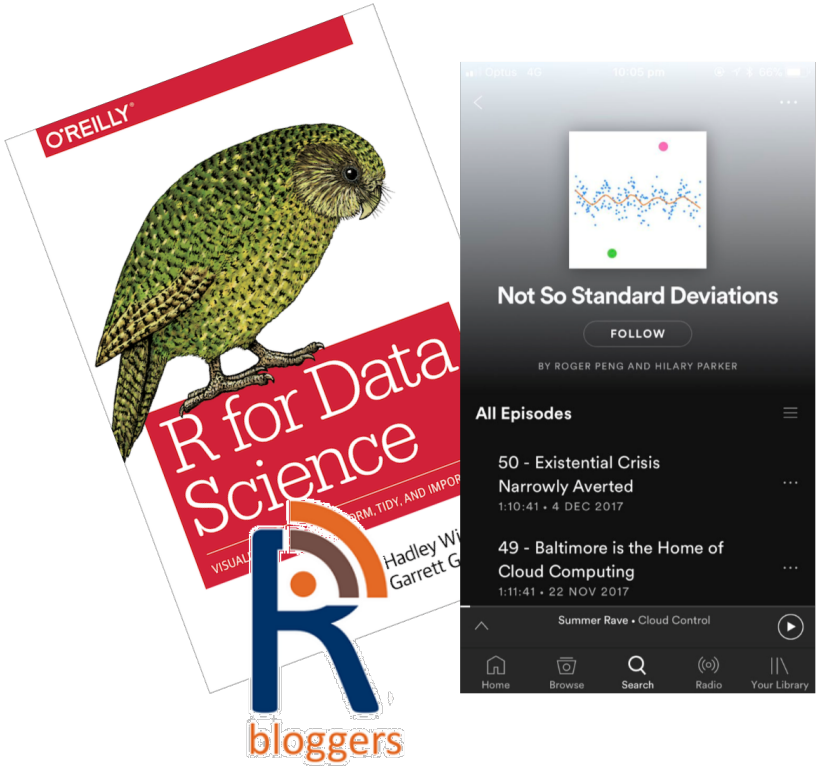
R Users & R ladies

The Community

R Users & R ladies

Twitter

The Community



R Users & R ladies

Twitter

Podcasts & Blogs

The Community

R Users & R ladies

Twitter

Podcasts & Blogs

Stack Overflow

Packages

The Tidyverse

The Tidyverse

simpler *

The Tidyverse

simpler *

human-readable

The Tidyverse

simpler *

human-readable

faster

The Tidyverse

simpler *

human-readable

faster

more intelligent

The Tidyverse

simpler *

human-readable

faster

more intelligent

magrittr's pipe

Import

Importing Data

Flat files, Data files, Databases, Web

Importing Data - CSV

```
library(tidyverse) #packages for data manipulation, exploration and visualization that
exchange_rates <- read.csv("data/ZAR_per_USD__Exchange_Rate_Detail_2000_to_2018-10.csv"

head(exchange_rates,3) # show top 3 rows of data frame
```

```
##         Date   Value
## 1 2018-10-01 14.1090
## 2 2018-09-28 14.1581
## 3 2018-09-27 14.1595
```

```
tail(exchange_rates,3) # show bottom 3 rows of data frame
```

```
##            Date  Value
## 4684 2000-01-06 6.0825
## 4685 2000-01-05 6.0881
## 4686 2000-01-04 6.1188
```

Importing Data - CSV

```
exchange_rates <- exchange_rates %>%
  filter(as.Date(exchange_rates$Date) >= as.Date('2018-01-01')) # filter for 2018 exch

head(exchange_rates,3)
```

```
##         Date   Value
## 1 2018-10-01 14.1090
## 2 2018-09-28 14.1581
## 3 2018-09-27 14.1595
```

```
tail(exchange_rates,3)
```

```
##            Date   Value
## 186 2018-01-04 12.3018
## 187 2018-01-03 12.3954
## 188 2018-01-02 12.3330
```

Importing Data - Excel Workbooks

```
library(readxl) # for reading excel files
path <- "data/gapminder_messy.xlsx"
combined_data <- excel_sheets(path) %>%
  map_df(~ {
    read_excel(path, sheet = .x,
               skip = 4, trim_ws = TRUE) %>%
      mutate(year = as.numeric(.x))
  }) %>%
  select(country, year, everything())
```

Importing Data - Excel Workbooks

```
head(combined_data,3)
```

```
## # A tibble: 3 x 5
##   country      year gdpPercap lifeExp      pop
##   <chr>       <dbl>     <dbl>   <dbl>    <dbl>
## 1 Afghanistan  2007      975.    43.8 31889923
## 2 Albania      2007     5937.    76.4  3600523
## 3 Algeria      2007     6223.    72.3 33333216
```

```
tail(combined_data,3)
```

```
## # A tibble: 3 x 5
##   country      year gdpPercap lifeExp     pop
##   <chr>       <dbl>     <dbl>   <dbl>   <dbl>
## 1 Yemen, Rep.  1952      782.    32.5 4963829
## 2 Zambia       1952     1147.    42.0 2672000
## 3 Zimbabwe     1952      407.    48.5 3080907
```

Importing Data - SPSS

```
library(haven) # for reading SPSS, SAS, STATA files
survey_data <- read_sav("data/sample_spss_data.sav")
head(survey_data)
```

```
## # A tibble: 6 x 682
##   SubsID responseid respid dCOUNTRY fQ1bGender foQ1aAge_1 dAge  Q2Kids_1
##   <chr>       <dbl>  <dbl> <dbl+lb> <dbl+lbl>       <dbl> <dbl>    <dbl>
## 1 21888~        345    347 1        1                  68 3            2
## 2 23586~        185    187 1        1                  51 2            2
## 3 23739~        108    110 1        2                  57 3            2
## 4 25015~        524    526 1        2                  53 2            2
## 5 25405~        349    351 1        2                  25 1            2
## 6 20004~        823    825 1        2                  54 2            1
## # ... with 674 more variables: Q2Kids_2 <dbl>, Q2Kids_3 <dbl>,
## #   dKIDS <dbl+lbl>, Q3 <dbl+lbl>, dNET <dbl+lbl>, fQ2UKRegion <dbl+lbl>,
## #   dRegionUK <dbl+lbl>, fQ2FRRegion <dbl+lbl>, dRegionFR <dbl+lbl>,
## #   fQ2DERegion <dbl+lbl>, dRegionDE <dbl+lbl>, Q2TURRegion <dbl+lbl>,
## #   Q2IncomeUK <dbl+lbl>, dIncomeUK <dbl+lbl>, Q2IncomeFR <dbl+lbl>,
## #   dIncomeFR <dbl+lbl>, Q2IncomeDE <dbl+lbl>, dIncomeDE <dbl+lbl>,
## #   Q2IncomeTUR <dbl+lbl>, dIncomeTUR <dbl+lbl>, Q4_1 <dbl+lbl>,
## #   Q4_2 <dbl+lbl>, Q4_3 <dbl+lbl>, Q4_4 <dbl+lbl>, Q4_5 <dbl+lbl>,
## #   Q4_6 <dbl+lbl>, Q4_99 <dbl+lbl>, Q6_1 <dbl+lbl>, Q6_2 <dbl+lbl>,
## #   Q6_3 <dbl+lbl>, Q6_4 <dbl+lbl>, Q6_5 <dbl+lbl>, Q6_98 <dbl+lbl>,
## #   Q6_99 <dbl+lbl>, dBiscLMAll_1 <dbl+lbl>, dBiscLMAll_2 <dbl+lbl>,
```

Importing Data - SPSS

```
# convert spss label codes into labels
labelled_survey_data <- as_factor(survey_data)
head(labelled_survey_data)
```

```
## # A tibble: 6 x 682
##   SubsID responseid respid dCOUNTRY fQ1bGender foQ1aAge_1 dAge  Q2Kids_1
##   <chr>       <dbl>  <dbl> <fct>    <fct>           <dbl> <fct>    <dbl>
## 1 21888~        345    347 UK       Male               68 55+          2
## 2 23586~        185    187 UK       Male               51 35-54        2
## 3 23739~        108    110 UK       Female             57 55+          2
## 4 25015~        524    526 UK       Female             53 35-54        2
## 5 25405~        349    351 UK       Female             25 16-34        2
## 6 20004~        823    825 UK       Female             54 35-54        1
## # ... with 674 more variables: Q2Kids_2 <dbl>, Q2Kids_3 <dbl>,
## #   dKIDS <fct>, Q3 <fct>, dNET <fct>, fQ2UKRegion <fct>, dRegionUK <fct>,
## #   fQ2FRRegion <fct>, dRegionFR <fct>, fQ2DERegion <fct>,
## #   dRegionDE <fct>, Q2TURRegion <fct>, Q2IncomeUK <fct>, dIncomeUK <fct>,
## #   Q2IncomeFR <fct>, dIncomeFR <fct>, Q2IncomeDE <fct>, dIncomeDE <fct>,
## #   Q2IncomeTUR <fct>, dIncomeTUR <fct>, Q4_1 <fct>, Q4_2 <fct>,
## #   Q4_3 <fct>, Q4_4 <fct>, Q4_5 <fct>, Q4_6 <fct>, Q4_99 <fct>,
## #   Q6_1 <fct>, Q6_2 <fct>, Q6_3 <fct>, Q6_4 <fct>, Q6_5 <fct>,
## #   Q6_98 <fct>, Q6_99 <fct>, dBiscLMAll_1 <fct>, dBiscLMAll_2 <fct>,
## #   dBiscLMAll_3 <fct>, dBiscLMAll_4 <fct>, dBiscLMAll_5 <fct>,
## #   dBiscLMAll_6 <fct>, dTypLMAll_1 <fct>, dTypLMAll_2 <fct>,
```

Importing Data - SPSS

```
labelled_survey_data %>%
count(fQ1bGender, dAge) %>%
spread(fQ1bGender, n)
```

```
## # A tibble: 3 x 3
##   dAge   Male Female
##   <fct> <int>  <int>
## 1 16-34   510    596
## 2 35-54   664    612
## 3 55+     362    480
```

Importing Data - Databases

```
library(DBI) # for connecting to databases

con <- dbConnect(
drv = RMySQL::MySQL(),
dbname = "shinydemo",
host = "shiny-demo.csa7qlmguqrf.us-east-1.rds.amazonaws.com",
username = "guest",
password = "guest"
)

dbListTables(con)
```

```
## [1] "City"            "Country"         "CountryLanguage"
```

Importing Data - Databases

```
cities <- tbl(con, "City")
cities
```

```
## # Source:   table<City> [?? x 5]
## # Database: mysql 10.0.17-MariaDB
## #   [guest@shiny-demo.csa7qlmguqrf.us-east-1.rds.amazonaws.com:/shinydemo]
##       ID Name           CountryCode District      Population
##    <dbl> <chr>          <chr>       <chr>              <dbl>
##  1     1 Kabul          AFG         Kabol            1780000
##  2     2 Qandahar       AFG         Qandahar          237500
##  3     3 Herat          AFG         Herat             186800
##  4     4 Mazar-e-Sharif AFG         Balkh             127800
##  5     5 Amsterdam      NLD         Noord-Holland     731200
##  6     6 Rotterdam      NLD         Zuid-Holland      593321
##  7     7 Haag           NLD         Zuid-Holland      440900
##  8     8 Utrecht        NLD         Utrecht           234323
##  9     9 Eindhoven      NLD         Noord-Brabant     201843
## 10    10 Tilburg        NLD         Noord-Brabant     193238
## # ... with more rows
```

Importing Data - Web APIs

```
library(jsonlite) #JSON parser and generator

github_issues <- function(user, repo){
  url <- paste0("https://api.github.com/repos/", user, "/", repo, "/issues")
  return(fromJSON(url))
}
dplyr_github_issues <- github_issues("hadley","dplyr")

head(dplyr_github_issues$title)
```

```
## [1] "Group_by for columns with dates as column name"
## [2] "Document `.data` and `!!` in manip verbs"
## [3] "Document semantics of mutate() and filter() on grouped tibbles"
## [4] "Review quasiquotation examples"
## [5] "Document semantics of colwise operations on grouping variables"
## [6] "Add group_cols() selection helper"
```

Tidy & Transform

Wait, what do you mean tidy?

Happy families are all alike;
every unhappy family is unhappy in its own way

Leo Tolstoy

Wait, what do you mean tidy?

Wait, what do you mean tidy?

	treatmenta	treatmentb
John Smith	—	2
Jane Doe	16	11
Mary Johnson	3	1

Wait, what do you mean tidy?

	treatmenta	treatmentb
John Smith	—	2
Jane Doe	16	11
Mary Johnson	3	1

	John Smith	Jane Doe	Mary Johnson
treatmenta	—	16	3
treatmentb	2	11	1

Wait, what do you mean tidy?

	treatmenta	treatmentb
John Smith	—	2
Jane Doe	16	11
Mary Johnson	3	1

	John Smith	Jane Doe	Mary Johnson
treatmenta	—	16	3
treatmentb	2	11	1

name	trt	result
John Smith	a	—
Jane Doe	a	16
Mary Johnson	a	3
John Smith	b	2
Jane Doe	b	11
Mary Johnson	b	1

Wait, what do you mean tidy?

name	trt	result
John Smith	a	—
Jane Doe	a	16
Mary Johnson	a	3
John Smith	b	2
Jane Doe	b	11
Mary Johnson	b	1

Wait, what do you mean tidy?

each variable forms a column

each observation forms a row

each type of observational unit forms a table

name	trt	result
John Smith	a	—
Jane Doe	a	16
Mary Johnson	a	3
John Smith	b	2
Jane Doe	b	11
Mary Johnson	b	1

Wait, what do you mean tidy?

each variable forms a column

each observation forms a row

each type of observational unit forms a table

name	trt	result
John Smith	a	—
Jane Doe	a	16
Mary Johnson	a	3
John Smith	b	2
Jane Doe	b	11
Mary Johnson	b	1

Wait, what do you mean tidy?

each variable forms a column

each observation forms a row

each type of observational unit forms a table

name	trt	result
John Smith	a	—
Jane Doe	a	16
Mary Johnson	a	3
John Smith	b	2
Jane Doe	b	11
Mary Johnson	b	1

Billboard Top 100 from 2000

```
head(raw_billboard)
```

```
## # A tibble: 6 x 10
##    year artist   track    time  genre date.entered   wk1   wk2   wk3   wk4
##   <int> <chr>    <chr>    <tim> <chr> <date>       <int> <int> <int> <int>
## 1  2000 Destiny~ Indepen~ 03:38 Rock  2000-09-23      78    63    49    33
## 2  2000 Santana  Maria, ~ 04:18 Rock  2000-02-12      15     8     6     5
## 3  2000 Madonna  Music    03:45 Rock  2000-08-12      41    23    18    14
## 4  2000 Aguiler~ Come On~ 03:38 Rock  2000-08-05      57    47    45    29
## 5  2000 Janet    Doesn't~ 04:17 Rock  2000-06-17      59    52    43    30
## 6  2000 Iglesia~ Be With~ 03:36 Latin 2000-04-01      63    45    34    23
```

So, how do we tidy our data?

`wide`

id	x	y	z
1	a	c	e
2	b	d	f

`long`

id	key	val
1	x	a
2	x	b
1	y	c
2	y	d
1	z	e
2	z	f

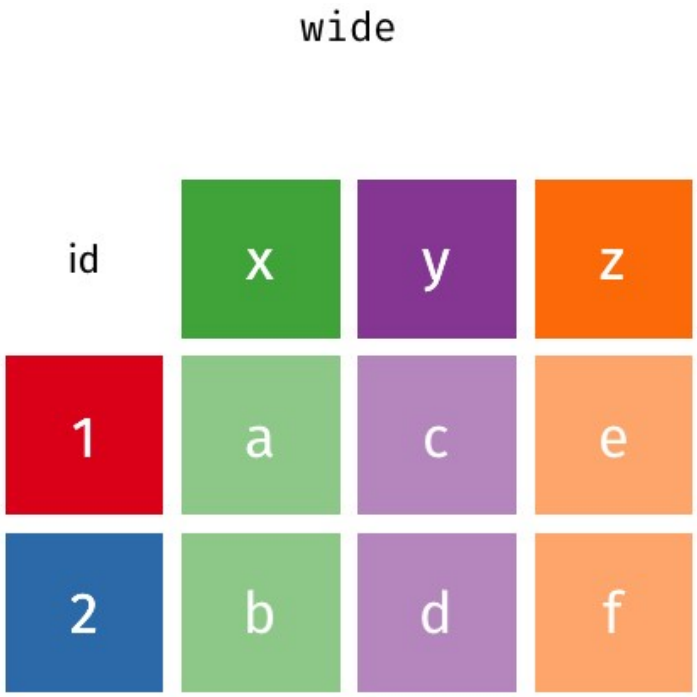
So, how do we tidy our data?

wide

id	x	y	z
1	a	c	e
2	b	d	f

long

id	key	val
1	x	a
2	x	b
1	y	c
2	y	d
1	z	e
2	z	f



Okay, let's try it out!

```
library(tidyverse)
raw_billboard %>% slice(1)
```

```
## # A tibble: 1 x 10
##    year artist   track    time  genre date.entered   wk1   wk2   wk3   wk4
##   <int> <chr>    <chr>    <tim> <chr> <date>       <int> <int> <int> <int>
## 1  2000 Destiny~ Indepen~ 03:38 Rock  2000-09-23      78    63    49    33
```

Okay, let's try it out!

```
library(tidyverse)
raw_billboard %>% slice(1)
```

```
## # A tibble: 1 x 10
##    year artist   track    time  genre date.entered   wk1   wk2   wk3   wk4
##   <int> <chr>    <chr>    <tim> <chr> <date>       <int> <int> <int> <int>
## 1  2000 Destiny~ Indepen~ 03:38 Rock  2000-09-23      78    63    49    33
```

```
raw_billboard %>% slice(1) %>%
  gather(key = "week", value = "rank", wk1, wk2, wk3, wk4)
```

```
## # A tibble: 4 x 8
##    year artist       track              time  genre date.entered week   rank
##   <int> <chr>        <chr>              <tim> <chr> <date>       <chr> <int>
## 1  2000 Destiny's C~ Independent Wom~ 03:38 Rock  2000-09-23   wk1      78
## 2  2000 Destiny's C~ Independent Wom~ 03:38 Rock  2000-09-23   wk2      63
## 3  2000 Destiny's C~ Independent Wom~ 03:38 Rock  2000-09-23   wk3      49
## 4  2000 Destiny's C~ Independent Wom~ 03:38 Rock  2000-09-23   wk4      33
```

Awesome, let's do it!

```
tidy_billboard <- raw_billboard %>%
                      gather(key = "week", value = "rank", wk1, wk2, wk3, wk4)
```

Awesome, let's do it!

```
tidy_billboard <- raw_billboard %>%
                        gather(key = "week", value = "rank", wk1, wk2, wk3, wk4)
head(tidy_billboard)
```

```
## # A tibble: 6 x 8
##    year artist      track              time  genre date.entered week   rank
##   <int> <chr>       <chr>              <tim> <chr> <date>       <chr> <int>
## 1  2000 Destiny's ~ Independent Wome~ 03:38 Rock  2000-09-23   wk1      78
## 2  2000 Santana     Maria, Maria       04:18 Rock  2000-02-12   wk1      15
## 3  2000 Madonna     Music              03:45 Rock  2000-08-12   wk1      41
## 4  2000 Aguilera, ~ Come On Over Bab~ 03:38 Rock  2000-08-05   wk1      57
## 5  2000 Janet       Doesn't Really M~ 04:17 Rock  2000-06-17   wk1      59
## 6  2000 Iglesias, ~ Be With You        03:36 Latin 2000-04-01   wk1      63
```

Transform - Arrange

```
tidy_billboard <- tidy_billboard %>%
                      arrange(artist, track)
```

Transform - Arrange

```
tidy_billboard <- tidy_billboard %>%
                        arrange(artist, track)
head(tidy_billboard)
```

```
## # A tibble: 6 x 8
##    year artist  track                    time  genre date.entered week   rank
##   <int> <chr>   <chr>                    <tim> <chr> <date>       <chr> <int>
## 1  2000 2 Pac   Baby Don't Cry (Keep~ 04:22 Rap   2000-02-26   wk1      87
## 2  2000 2 Pac   Baby Don't Cry (Keep~ 04:22 Rap   2000-02-26   wk2      82
## 3  2000 2 Pac   Baby Don't Cry (Keep~ 04:22 Rap   2000-02-26   wk3      72
## 4  2000 2 Pac   Baby Don't Cry (Keep~ 04:22 Rap   2000-02-26   wk4      77
## 5  2000 2Ge+her The Hardest Part Of ~ 03:15 R&B   2000-09-02   wk1      91
## 6  2000 2Ge+her The Hardest Part Of ~ 03:15 R&B   2000-09-02   wk2      87
```

Transform - Mutate

```
library(lubridate)
tidy_billboard_pop <- tidy_billboard %>%
                          mutate(month.entered = month(tidy_billboard$date.entered, 
```

Transform - Mutate

```
library(lubridate)
tidy_billboard_pop <- tidy_billboard %>%
                            mutate(month.entered = month(tidy_billboard$date.entered, 
head(tidy_billboard_pop)
```

```
## # A tibble: 6 x 9
##    year artist  track   time  genre date.entered week   rank month.entered
##   <int> <chr>   <chr>   <tim> <chr> <date>       <chr> <int> <ord>
## 1  2000 2 Pac   Baby D~ 04:22 Rap   2000-02-26   wk1      87 Feb
## 2  2000 2 Pac   Baby D~ 04:22 Rap   2000-02-26   wk2      82 Feb
## 3  2000 2 Pac   Baby D~ 04:22 Rap   2000-02-26   wk3      72 Feb
## 4  2000 2 Pac   Baby D~ 04:22 Rap   2000-02-26   wk4      77 Feb
## 5  2000 2Ge+her The Ha~ 03:15 R&B   2000-09-02   wk1      91 Sep
## 6  2000 2Ge+her The Ha~ 03:15 R&B   2000-09-02   wk2      87 Sep
```

Transform - Filter

```
tidy_billboard_pop <- tidy_billboard %>%
                          filter(genre == "Pop" & month(date.entered) < 7)
```

Transform - Filter

```
tidy_billboard_pop <- tidy_billboard %>%
                          filter(genre == "Pop" & month(date.entered) < 7)
head(tidy_billboard_pop)
```

```
## # A tibble: 6 x 8
##    year artist       track              time  genre date.entered week   rank
##   <int> <chr>        <chr>              <tim> <chr> <date>       <chr> <int>
## 1  2000 Anastacia    I'm Outta Love     04:01 Pop   2000-04-01   wk1      92
## 2  2000 Anastacia    I'm Outta Love     04:01 Pop   2000-04-01   wk2      NA
## 3  2000 Anastacia    I'm Outta Love     04:01 Pop   2000-04-01   wk3      NA
## 4  2000 Anastacia    I'm Outta Love     04:01 Pop   2000-04-01   wk4      95
## 5  2000 Fabian, Lara I Will Love Aga~ 03:43 Pop   2000-06-10   wk1      91
## 6  2000 Fabian, Lara I Will Love Aga~ 03:43 Pop   2000-06-10   wk2      80
```

Transform - Select

```
tidy_billboard_pop <- tidy_billboard_pop %>%
                          select(-genre)
```

Transform - Select

```
tidy_billboard_pop <- tidy_billboard_pop %>%
                            select(-genre)
head(tidy_billboard_pop)
```

```
## # A tibble: 6 x 7
##    year artist       track              time   date.entered week   rank
##   <int> <chr>        <chr>              <time> <date>       <chr> <int>
## 1  2000 Anastacia    I'm Outta Love     04:01  2000-04-01   wk1      92
## 2  2000 Anastacia    I'm Outta Love     04:01  2000-04-01   wk2      NA
## 3  2000 Anastacia    I'm Outta Love     04:01  2000-04-01   wk3      NA
## 4  2000 Anastacia    I'm Outta Love     04:01  2000-04-01   wk4      95
## 5  2000 Fabian, Lara I Will Love Again 03:43  2000-06-10   wk1      91
## 6  2000 Fabian, Lara I Will Love Again 03:43  2000-06-10   wk2      80
```

Transform - Summarise

```
  tidy_billboard_pop %>%
    summarise(mean.rank = mean(rank, na.rm = T))
```

```
## # A tibble: 1 x 1
##   mean.rank
##       <dbl>
## 1      71.7
```

Transform - Group By & Summarise

```
  tidy_billboard_pop %>%
      group_by(artist) %>%
      summarise(mean.rank = mean(rank, na.rm = T))
```

```
## # A tibble: 5 x 2
##   artist       mean.rank
##   <chr>            <dbl>
## 1 Anastacia         93.5
## 2 Fabian, Lara      76.8
## 3 Hoku              50.2
## 4 M2M               89.8
## 5 Moore, Mandy      59
```

Transform

`tidy_billboard`

Transform

```
tidy_billboard %>%
    filter(genre == "Pop" & month(date.entered) < 7)
```

Transform

```
tidy_billboard %>%
    filter(genre == "Pop" & month(date.entered) < 7) %>%
    select(-genre)
```

Transform

```
tidy_billboard %>%
    filter(genre == "Pop" & month(date.entered) < 7) %>%
    select(-genre) %>%
    group_by(artist)
```

Transform

```
tidy_billboard %>%
    filter(genre == "Pop" & month(date.entered) < 7) %>%
    select(-genre) %>%
    group_by(artist) %>%
    summarise(mean.rank = mean(rank, na.rm = T))
```

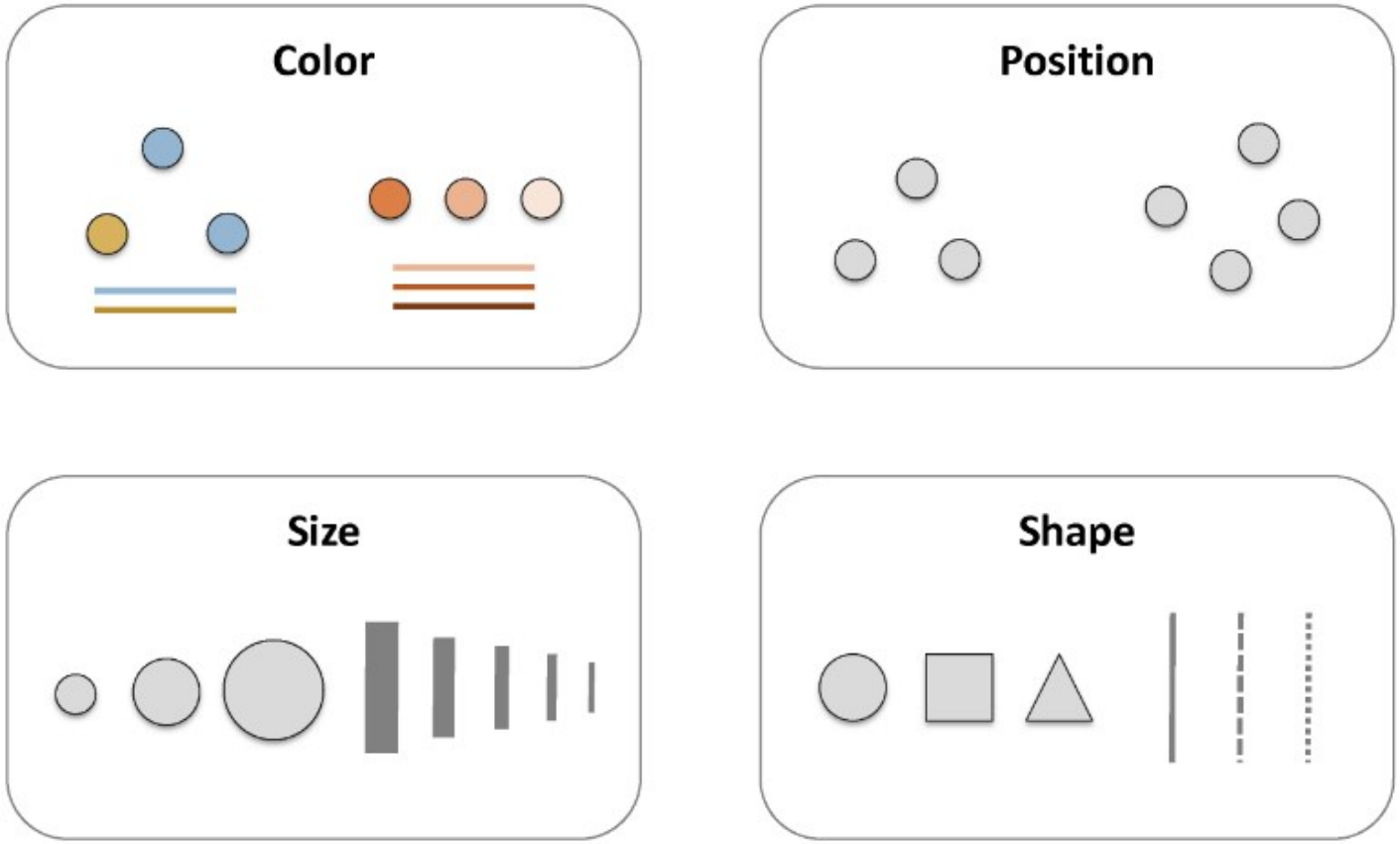
Transform

```
  tidy_billboard %>%
      filter(genre == "Pop" & month(date.entered) < 7) %>%
      select(-genre) %>%
      group_by(artist) %>%
      summarise(mean.rank = mean(rank, na.rm = T))
```

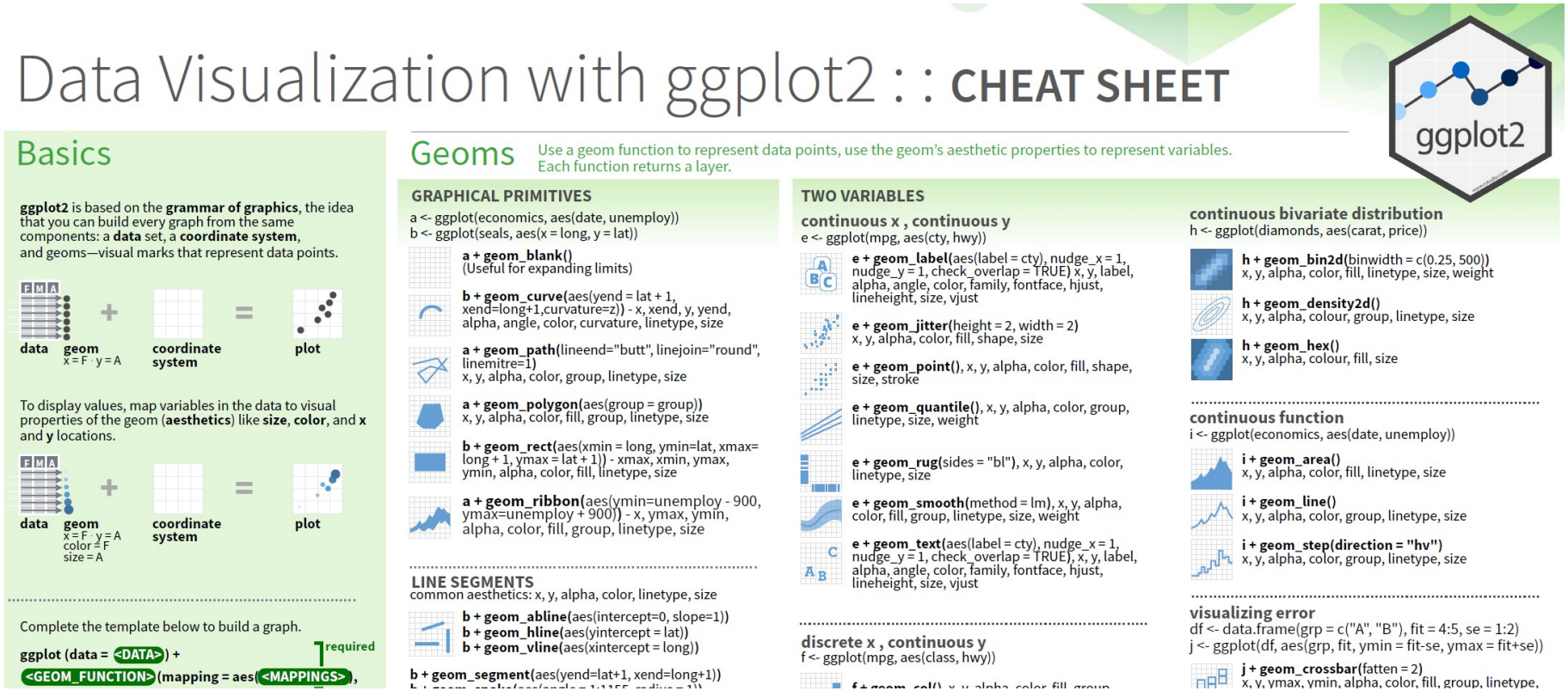
```
## # A tibble: 5 x 2
##   artist       mean.rank
##   <chr>            <dbl>
## 1 Anastacia         93.5
## 2 Fabian, Lara      76.8
## 3 Hoku              50.2
## 4 M2M               89.8
## 5 Moore, Mandy      59
```

Visualise

Data Visualisation



ggplot2



Data Visualization with ggplot2 : : CHEAT SHEET

Basics

ggplot2 is based on the **grammar of graphics**, the idea that you can build every graph from the same components: a **data** set, a **coordinate system**, and geoms—visual marks that represent data points.

data + geom (x = F, y = A) = coordinate system → plot

To display values, map variables in the data to visual properties of the geom (**aesthetics**) like **size**, **color**, and **x** and **y** locations.

data + geom (x = F, y = A, color = F, size = A) = coordinate system → plot

Complete the template below to build a graph.

```
ggplot (data = <DATA>) +                    required
<GEOM_FUNCTION> (mapping = aes( <MAPPINGS> ),
```

Geoms

Use a geom function to represent data points, use the geom's aesthetic properties to represent variables. Each function returns a layer.

GRAPHICAL PRIMITIVES

```
a <- ggplot(economics, aes(date, unemploy))
b <- ggplot(seals, aes(x = long, y = lat))
```

a + geom_blank()
(Useful for expanding limits)

b + geom_curve(aes(yend = lat + 1, xend=long+1,curvature=z)) - x, xend, y, yend, alpha, angle, color, curvature, linetype, size

a + geom_path(lineend="butt", linejoin="round", linemitre=1)
x, y, alpha, color, group, linetype, size

a + geom_polygon(aes(group = group))
x, y, alpha, color, fill, group, linetype, size

b + geom_rect(aes(xmin = long, ymin=lat, xmax= long + 1, ymax = lat + 1)) - xmax, xmin, ymax, ymin, alpha, color, fill, linetype, size

a + geom_ribbon(aes(ymin=unemploy - 900, ymax=unemploy + 900)) - x, ymax, ymin, alpha, color, fill, group, linetype, size

LINE SEGMENTS

common aesthetics: x, y, alpha, color, linetype, size

b + geom_abline(aes(intercept=0, slope=1))
b + geom_hline(aes(yintercept = lat))
b + geom_vline(aes(xintercept = long))

b + geom_segment(aes(yend=lat+1, xend=long+1))

TWO VARIABLES

continuous x , continuous y

```
e <- ggplot(mpg, aes(cty, hwy))
```

e + geom_label(aes(label = cty), nudge_x = 1, nudge_y = 1, check_overlap = TRUE) x, y, label, alpha, angle, color, family, fontface, hjust, lineheight, size, vjust

e + geom_jitter(height = 2, width = 2)
x, y, alpha, color, fill, shape, size

e + geom_point(), x, y, alpha, color, fill, shape, size, stroke

e + geom_quantile(), x, y, alpha, color, group, linetype, size, weight

e + geom_rug(sides = "bl"), x, y, alpha, color, linetype, size

e + geom_smooth(method = lm), x, y, alpha, color, fill, group, linetype, size, weight

e + geom_text(aes(label = cty), nudge_x = 1, nudge_y = 1, check_overlap = TRUE), x, y, label, alpha, angle, color, family, fontface, hjust, lineheight, size, vjust

discrete x , continuous y

```
f <- ggplot(mpg, aes(class, hwy))
```

continuous bivariate distribution

```
h <- ggplot(diamonds, aes(carat, price))
```

h + geom_bin2d(binwidth = c(0.25, 500))
x, y, alpha, color, fill, linetype, size, weight

h + geom_density2d()
x, y, alpha, colour, group, linetype, size

h + geom_hex()
x, y, alpha, colour, fill, size

continuous function

```
i <- ggplot(economics, aes(date, unemploy))
```

i + geom_area()
x, y, alpha, color, fill, linetype, size

i + geom_line()
x, y, alpha, color, group, linetype, size

i + geom_step(direction = "hv")
x, y, alpha, color, group, linetype, size

visualizing error

```
df <- data.frame(grp = c("A", "B"), fit = 4:5, se = 1:2)
j <- ggplot(df, aes(grp, fit, ymin = fit-se, ymax = fit+se))
```

j + geom_crossbar(fatten = 2)
x, y, ymax, ymin, alpha, color, fill, group, linetype,

RStudio Cheat Sheets - https://www.rstudio.com/resources/cheatsheets/

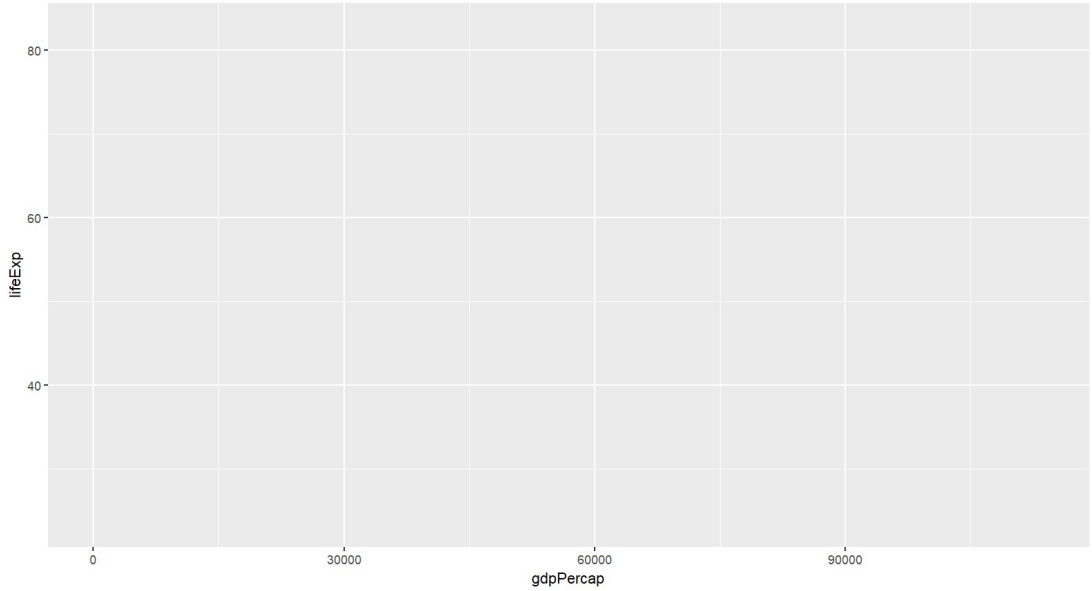
ggplot2

```
ggplot(gapminder)
```


This example inspired by **Gina Reynolds**, **Garrick Aden-Buie** and **Emi Tanaka**.

ggplot2

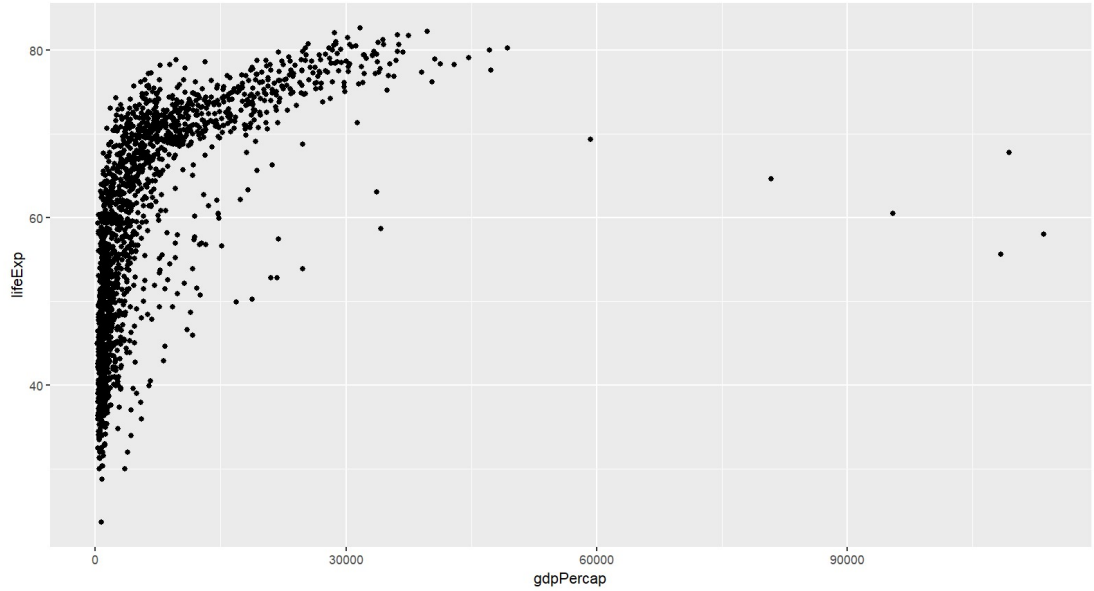
```
ggplot(gapminder) +
  aes(x = gdpPercap, y = lifeExp)
```



Note: `aes` normally sits within `ggplot` function!

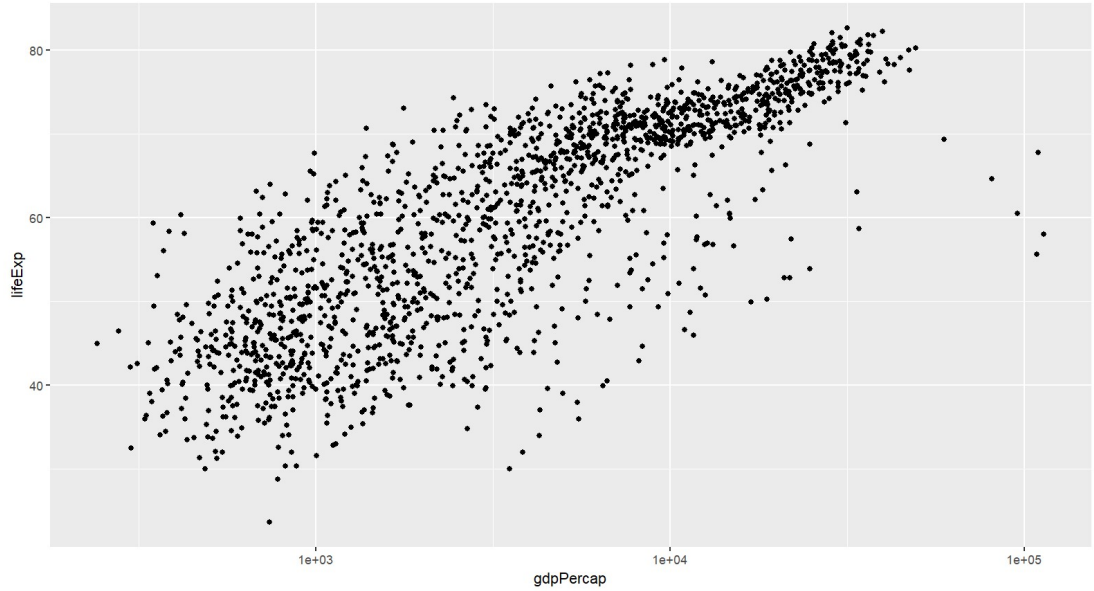
ggplot2

```
ggplot(gapminder) +
  aes(x = gdpPercap, y = lifeExp)
  geom_point()
```



ggplot2

```
ggplot(gapminder) +
  aes(x = gdpPercap, y = lifeExp)
  geom_point() +
  scale_x_log10()
```



ggplot2

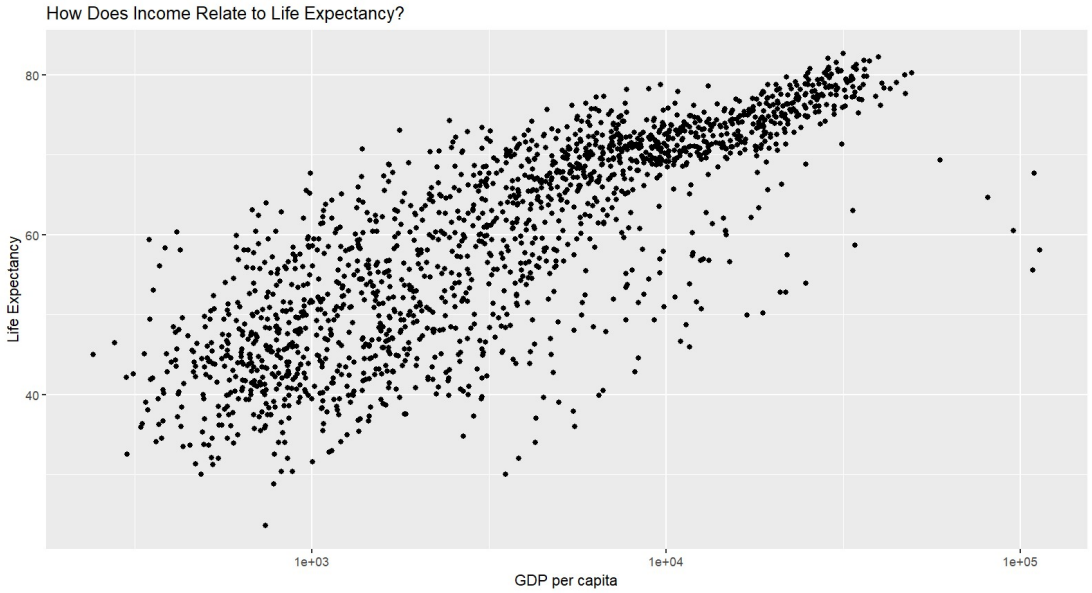
```
ggplot(gapminder) +
  aes(x = gdpPercap, y = lifeExp)
  geom_point() +
  scale_x_log10() +
  labs(x = "GDP per capita")
```

ggplot2

```
ggplot(gapminder) +
  aes(x = gdpPercap, y = lifeExp)
  geom_point() +
  scale_x_log10() +
  labs(x = "GDP per capita") +
  labs(y = "Life Expectancy")
```

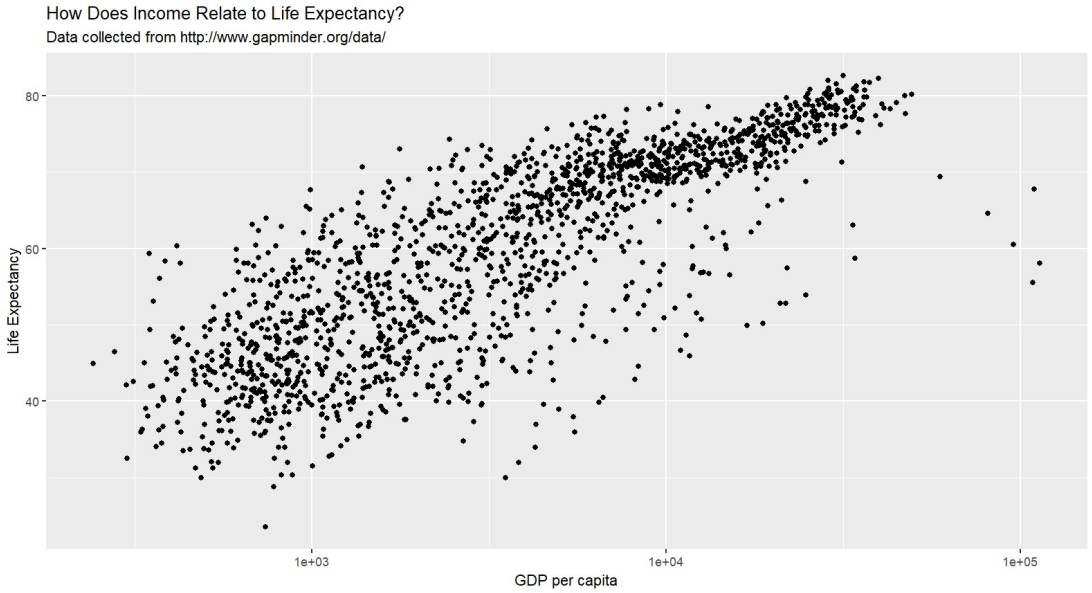
ggplot2

```
ggplot(gapminder) +
  aes(x = gdpPercap, y = lifeExp)
  geom_point() +
  scale_x_log10() +
  labs(x = "GDP per capita") +
  labs(y = "Life Expectancy") +
  labs(title="How Does Income Rela
```



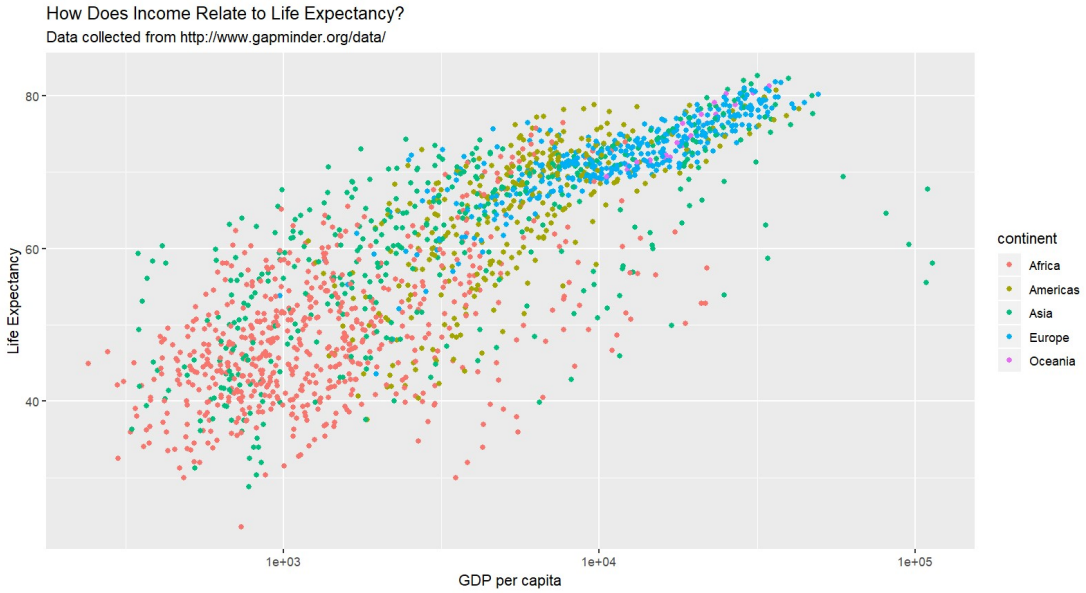
ggplot2

```
ggplot(gapminder) +
  aes(x = gdpPercap, y = lifeExp)
  geom_point() +
  scale_x_log10() +
  labs(x = "GDP per capita") +
  labs(y = "Life Expectancy") +
  labs(title="How Does Income Rela
  labs(subtitle="Data collected fr
```



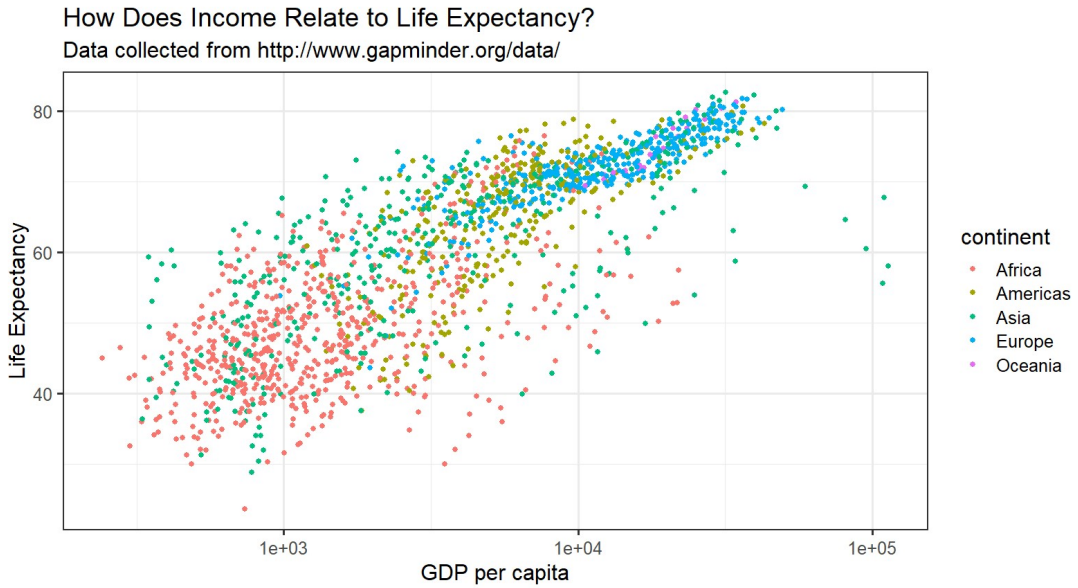
ggplot2

```
ggplot(gapminder) +
  aes(x = gdpPercap, y = lifeExp)
  geom_point() +
  scale_x_log10() +
  labs(x = "GDP per capita") +
  labs(y = "Life Expectancy") +
  labs(title="How Does Income Rela
  labs(subtitle="Data collected fr
  aes(color = continent)
```



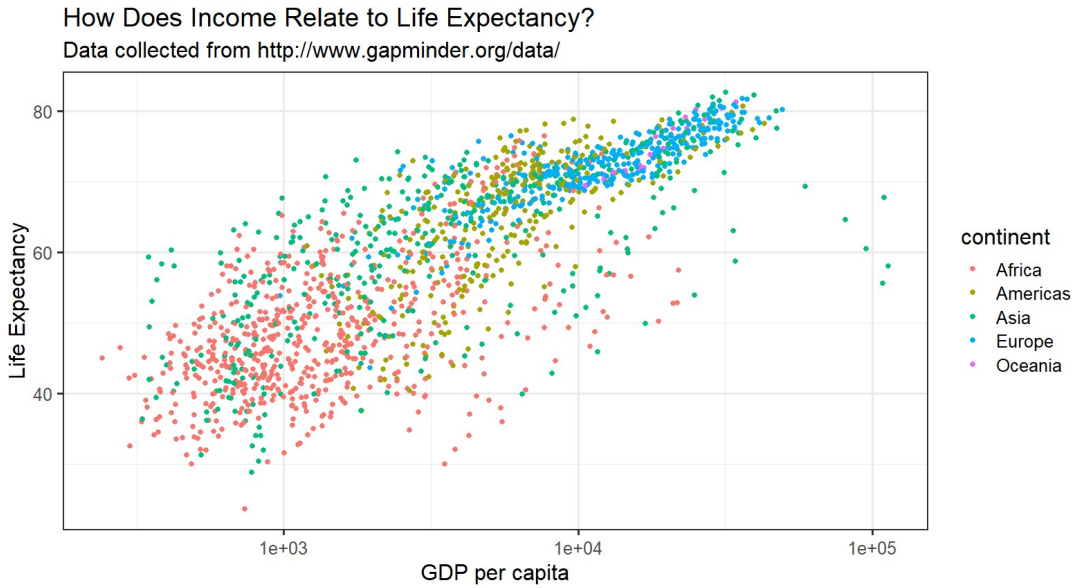
ggplot2

```
ggplot(gapminder) +
  aes(x = gdpPercap, y = lifeExp)
  geom_point() +
  scale_x_log10() +
  labs(x = "GDP per capita") +
  labs(y = "Life Expectancy") +
  labs(title="How Does Income Rela
  labs(subtitle="Data collected fr
  aes(color = continent) +
  theme_bw(base_size=16)
```



ggplot2

```
ggplot(gapminder) +
  aes(x = gdpPercap, y = lifeExp)
  geom_point() +
  scale_x_log10() +
  labs(x = "GDP per capita") +
  labs(y = "Life Expectancy") +
  labs(title="How Does Income Rela
  labs(subtitle="Data collected fr
  aes(color = continent) +
  theme_bw(base_size=16)
```



Book recommendation

Factfulness by Hans Rosling

Import + Transform + Visualise

```
library(sf)
library(ggplot2)
library(dplyr)
geo <- st_read("data/LocalMunicipa
head(geo)
```

```
## Simple feature collection with 6 features and 12 f
##          ## geometry type:  MULTIPOLYGON
##               ## dimension:       XY
## bbox:              xmin: 23.22161 ymin: -33.75659 xma
##              ## epsg (SRID):    4326
## proj4string:    +proj=longlat +datum=WGS84 +no_def
##   PROVINCE CATEGORY                          CAT2 CAT_
## 1       EC        A Metropolitan Municipality   BU
## 2       EC        B        Local Municipality EC10
## 3       EC        B        Local Municipality EC10
## 4       EC        B        Local Municipality EC10
## 5       EC        B        Local Municipality EC10
## 6       EC        B        Local Municipality EC10
##                                  MAP_TITLE DISTRICT
## 1 Buffalo City Metropolitan Municipality      BUF
## 2            Camdeboo Local Municipality     DC10
## 3    Blue Crane Route Local Municipality     DC10
## 4              Ikwezi Local Municipality     DC10
## 5              Makana Local Municipality     DC10
## 6             Ndlambe Local Municipality     DC10
##   Shape_Area MapNo  MUN_CD                          g
## 1  0.2445463     1   (BUF) MULTIPOLYGON (((27.917
## 2  1.1903525     4 (EC101) MULTIPOLYGON (((24.439
## 3  1.0653265     5 (EC102) MULTIPOLYGON (((26.165
## 4  0.4402566     6 (EC103) MULTIPOLYGON (((25.051
```

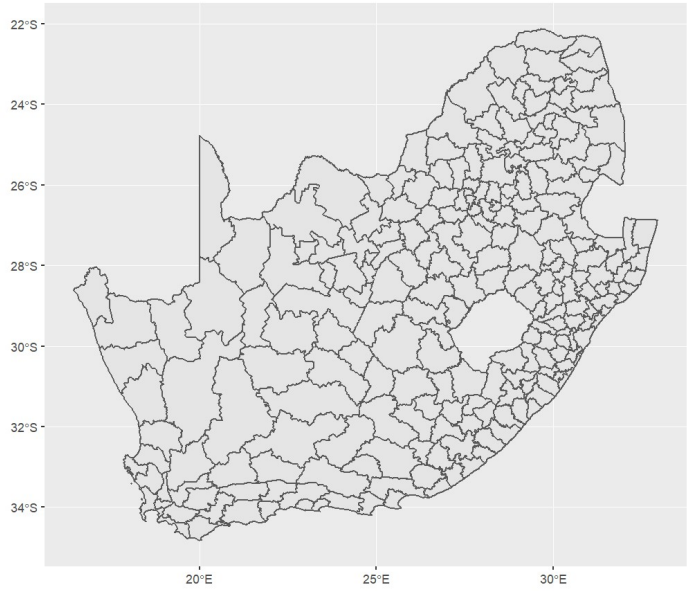
Import + Transform + Visualise

```
library(sf)
library(ggplot2)
library(dplyr)
geo <- st_read("data/LocalMunicipa
#head(geo)
largestMunicipalities <- geo %>%
  select(AREA, MUNICNAME) %>%
  arrange(desc(AREA)) %>%
  slice(1:10)
largestMunicipalities
```

```
## Simple feature collection with 10 features and 2 f
##          geometry type:  MULTIPOLYGON
##               dimension:       XY
## bbox:           xmin: 16.45189 ymin: -34.83417 xma
##              epsg (SRID):    4326
## proj4string:    +proj=longlat +datum=WGS84 +no_def
##       AREA        MUNICNAME                          g
## 1  36128.1           Hantam MULTIPOLYGON (((19.0130
## 2  32273.9  Karoo Hoogland MULTIPOLYGON (((20.5915
## 3  26358.0      Kai !Garib MULTIPOLYGON (((19.9991
## 4  23827.3 Kagisano/Molopo MULTIPOLYGON (((24.2660
## 5  22468.4            Mier MULTIPOLYGON (((20.8419
## 6  21916.6   Beaufort West MULTIPOLYGON (((23.6365
## 7  21779.8    //Khara Hais MULTIPOLYGON (((21.7100
## 8  20389.2          Ubuntu MULTIPOLYGON (((23.2229
## 9  20172.0    Joe Morolong MULTIPOLYGON (((23.9173
## 10 18332.8     Tsantsabane MULTIPOLYGON (((22.9624
```

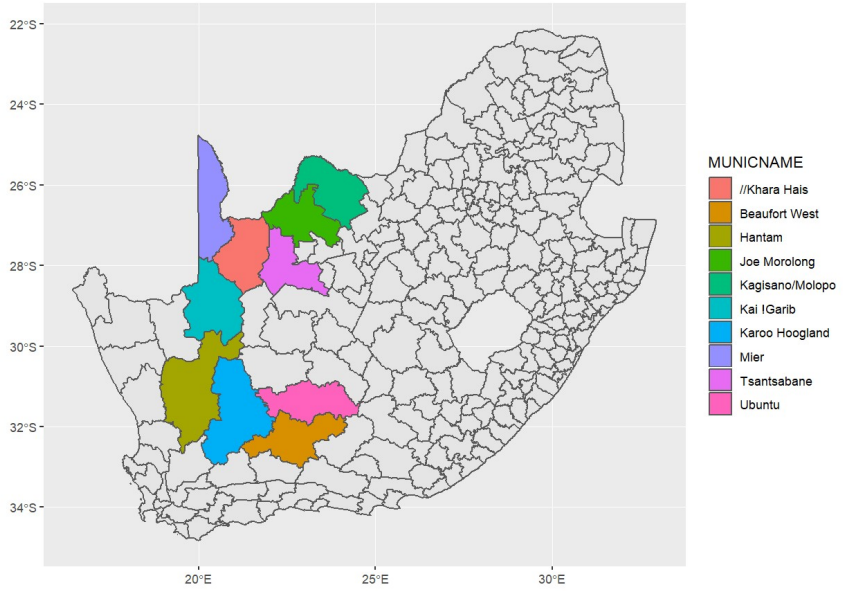
Import + Transform + Visualise

```
library(sf)
library(ggplot2)
library(dplyr)
geo <- st_read("data/LocalMunicipa
#head(geo)
largestMunicipalities <- geo %>%
  select(AREA, MUNICNAME) %>%
  arrange(desc(AREA)) %>%
  slice(1:10)
#largestMunicipalities
ggplot() +
  geom_sf(data = geo)
```



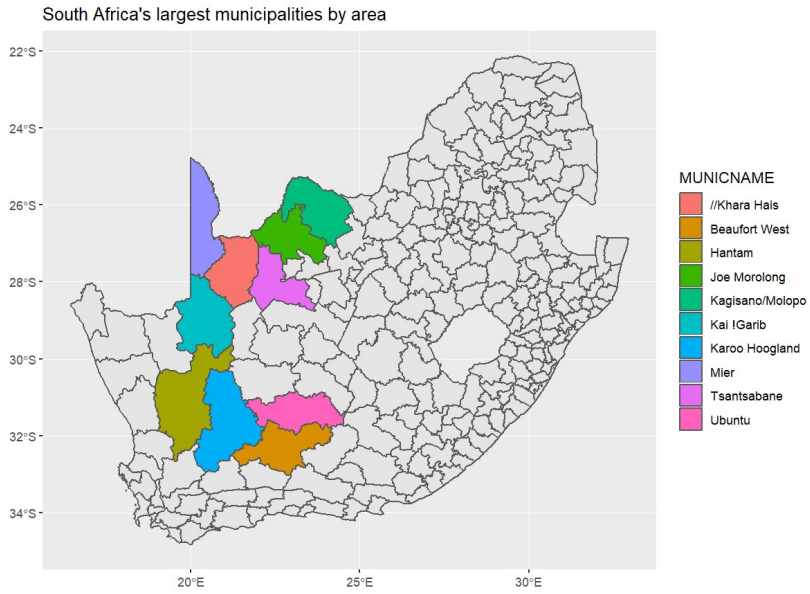
Import + Transform + Visualise

```
library(sf)
library(ggplot2)
library(dplyr)
geo <- st_read("data/LocalMunicipa
#head(geo)
largestMunicipalities <- geo %>%
  select(AREA, MUNICNAME) %>%
  arrange(desc(AREA)) %>%
  slice(1:10)
#largestMunicipalities
ggplot() +
  geom_sf(data = geo) +
  geom_sf(data = largestMunicipal
```



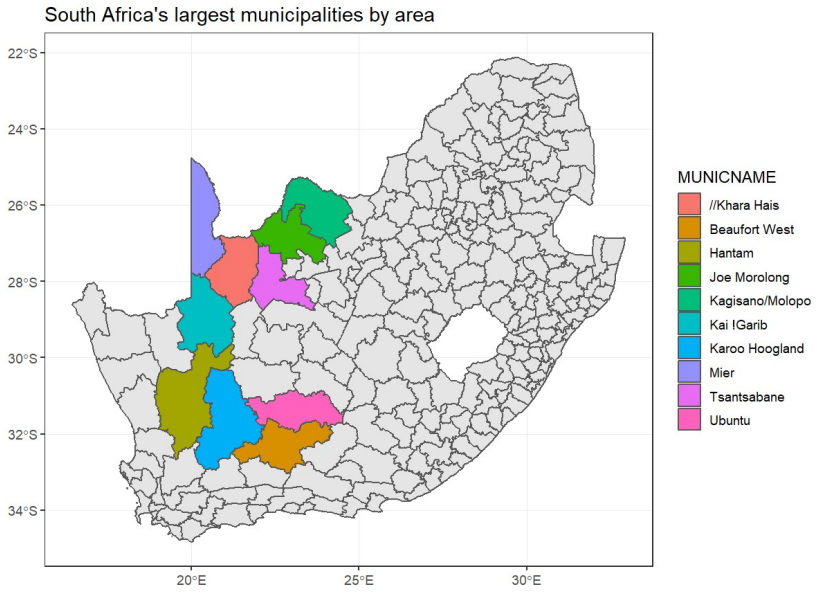
Import + Transform + Visualise

```
library(sf)
library(ggplot2)
library(dplyr)
geo <- st_read("data/LocalMunicipa
#head(geo)
largestMunicipalities <- geo %>%
  select(AREA, MUNICNAME) %>%
  arrange(desc(AREA)) %>%
  slice(1:10)
#largestMunicipalities
ggplot() +
  geom_sf(data = geo) +
  geom_sf(data = largestMunicipal
  labs(title= "South Africa's larg
```



Import + Transform + Visualise

```
library(sf)
library(ggplot2)
library(dplyr)
geo <- st_read("data/LocalMunicipa
#head(geo)
largestMunicipalities <- geo %>%
  select(AREA, MUNICNAME) %>%
  arrange(desc(AREA)) %>%
  slice(1:10)
#largestMunicipalities
ggplot() +
  geom_sf(data = geo) +
  geom_sf(data = largestMunicipal
  labs(title= "South Africa's larg
  theme_bw(base_size=12)
```



Model

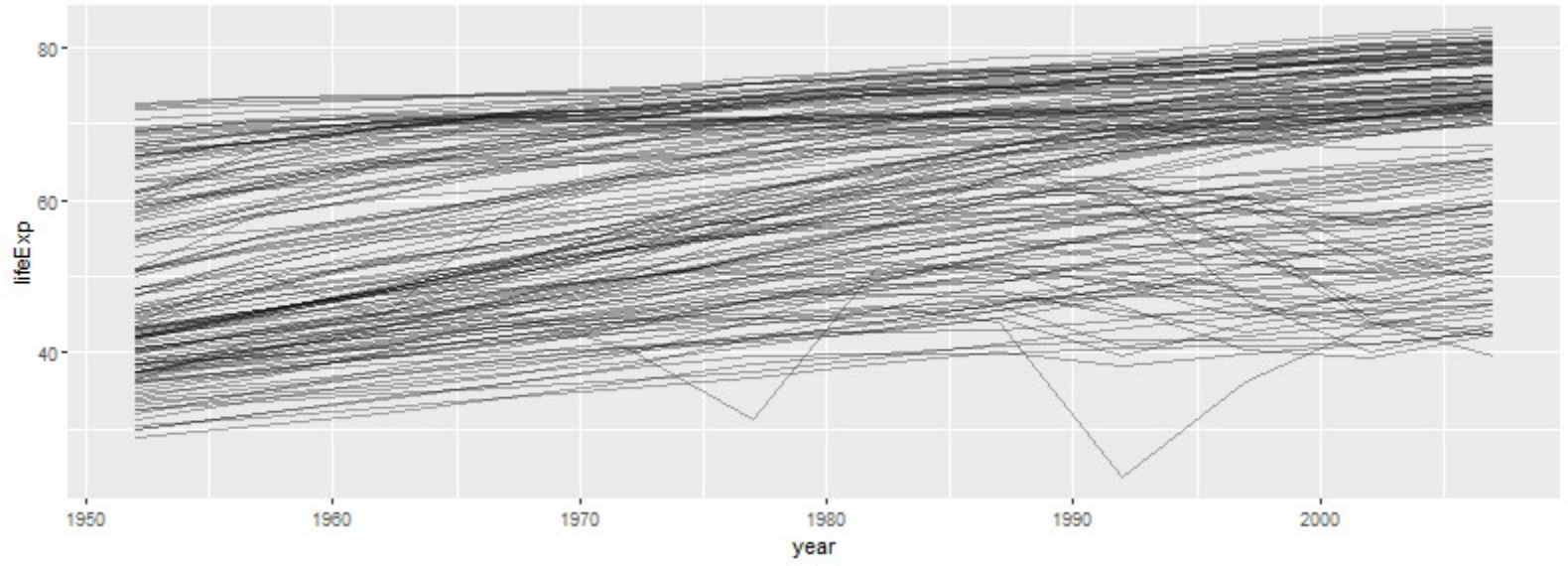
Transform + Visualise + Model

All models are wrong, but some are useful.

George Box

Model - Life Expectancy

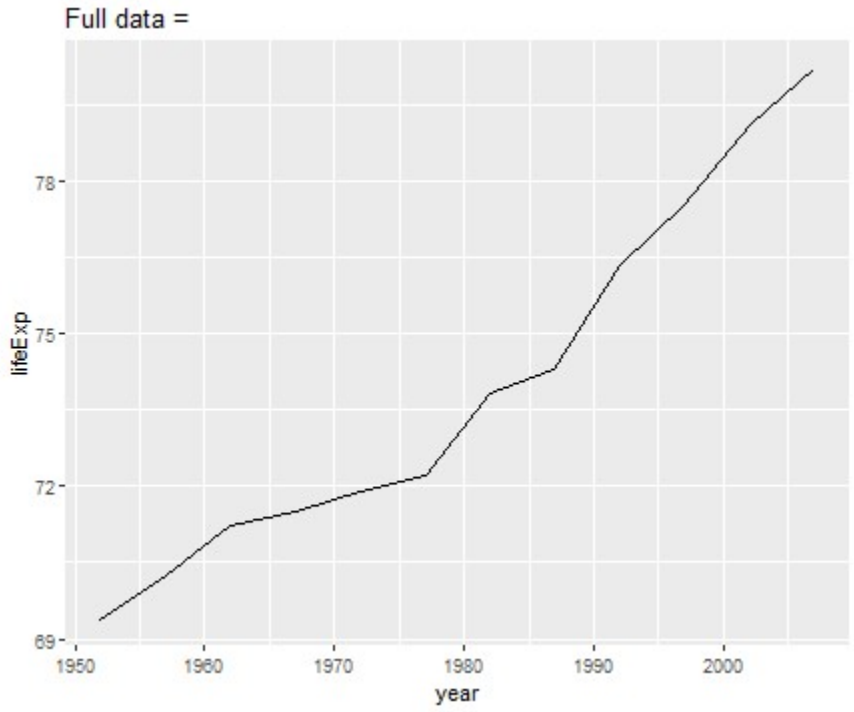
```
gapminder %>% ggplot(aes(year, lifeExp, group = country)) + geom_line(alpha = 1/3)
```



Model - New Zealand

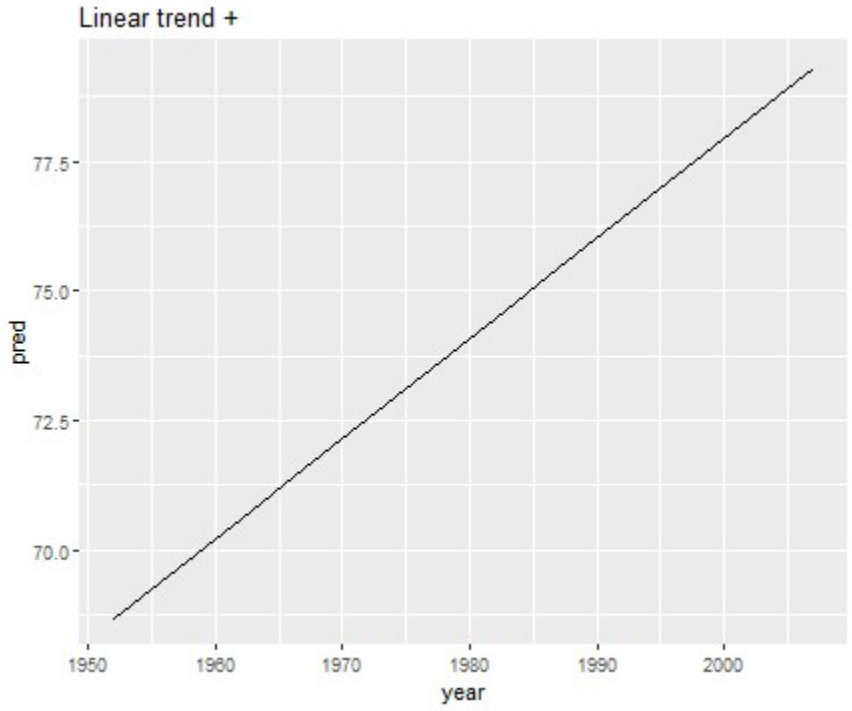
```
nz <- filter(gapminder, country=="New Zeal
```

```
nz %>%
  ggplot(aes(year, lifeExp)) +
  geom_line() +
  ggtitle("Full data = ")
```



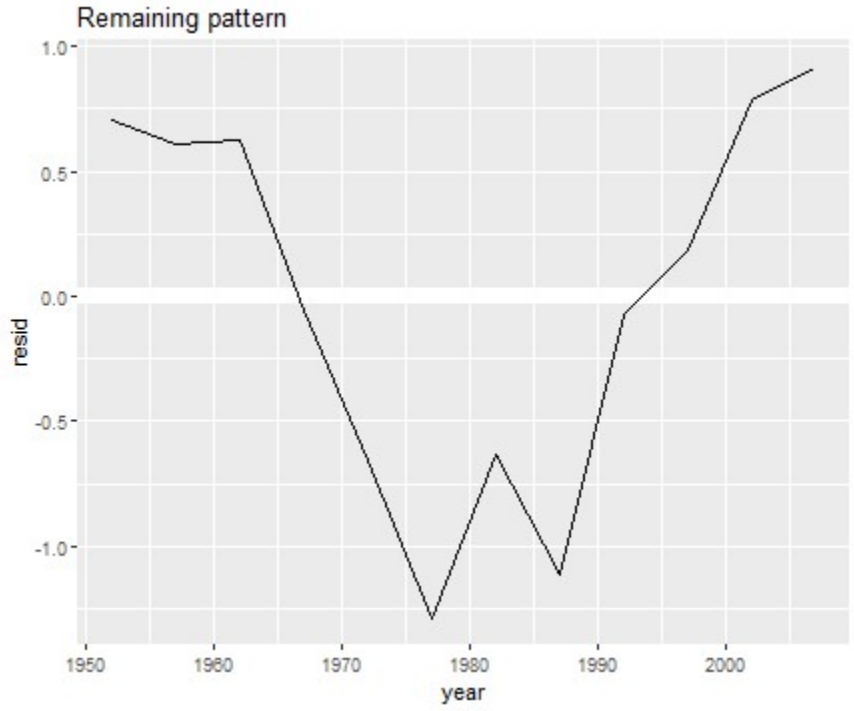
Model - New Zealand

```
library(modelr)

nz_mod <- lm(lifeExp ~ year, data = nz)

nz %>%
  add_predictions(nz_mod) %>%
  ggplot(aes(year, pred)) +
  geom_line() +
  ggtitle("Linear trend + ")
```



Model - New Zealand

```
nz %>%
  add_residuals(nz_mod) %>%
  ggplot(aes(year, resid)) +
  geom_hline(yintercept = 0,
             colour = "white",
             size = 3) +
  geom_line() +
  ggtitle("Remaining pattern")
```



Model - Set Up

```
by_country <- gapminder %>%
                group_by(country, continent) %>%
                nest()

head(by_country)
```

```
## # A tibble: 6 x 3
##   country     continent data
##   <fct>       <fct>     <list>
## 1 Afghanistan Asia      <tibble [12 x 4]>
## 2 Albania     Europe    <tibble [12 x 4]>
## 3 Algeria     Africa    <tibble [12 x 4]>
## 4 Angola      Africa    <tibble [12 x 4]>
## 5 Argentina   Americas  <tibble [12 x 4]>
## 6 Australia   Oceania   <tibble [12 x 4]>
```

Data Frame Inception

```
by_country$data[[1]]
```

```
## # A tibble: 12 x 4
##     year lifeExp      pop gdpPercap
##    <int>   <dbl>    <int>     <dbl>
##  1  1952    28.8  8425333      779.
##  2  1957    30.3  9240934      821.
##  3  1962    32.0 10267083      853.
##  4  1967    34.0 11537966      836.
##  5  1972    36.1 13079460      740.
##  6  1977    38.4 14880372      786.
##  7  1982    39.9 12881816      978.
##  8  1987    40.8 13867957      852.
##  9  1992    41.7 16317921      649.
## 10  1997    41.8 22227415      635.
## 11  2002    42.1 25268405      727.
## 12  2007    43.8 31889923      975.
```

Model - Life Expectancy

```
country_model <- function(df) {
  lm(lifeExp ~ year, data = df)
}
```

Model - Life Expectancy

```
country_model <- function(df) {
  lm(lifeExp ~ year, data = df)
}

by_country <- by_country %>%
  mutate(model = map(data, country_model))
head(by_country)
```

```
## # A tibble: 6 x 4
##   country     continent data              model
##   <fct>       <fct>     <list>            <list>
## 1 Afghanistan Asia      <tibble [12 x 4]> <S3: lm>
## 2 Albania     Europe    <tibble [12 x 4]> <S3: lm>
## 3 Algeria     Africa    <tibble [12 x 4]> <S3: lm>
## 4 Angola      Africa    <tibble [12 x 4]> <S3: lm>
## 5 Argentina   Americas  <tibble [12 x 4]> <S3: lm>
## 6 Australia   Oceania   <tibble [12 x 4]> <S3: lm>
```

Enjoying that Data Frame Goodness

```
by_country %>%
  filter(continent == "Europe") %>%
  head()
```

```
## # A tibble: 6 x 4
##   country                continent data              model
##   <fct>                  <fct>     <list>            <list>
## 1 Albania                Europe    <tibble [12 x 4]> <S3: lm>
## 2 Austria                Europe    <tibble [12 x 4]> <S3: lm>
## 3 Belgium                Europe    <tibble [12 x 4]> <S3: lm>
## 4 Bosnia and Herzegovina Europe    <tibble [12 x 4]> <S3: lm>
## 5 Bulgaria               Europe    <tibble [12 x 4]> <S3: lm>
## 6 Croatia                Europe    <tibble [12 x 4]> <S3: lm>
```

Enjoying that Data Frame Goodness

```
by_country %>%
  arrange(continent, country) %>%
  head()
```

```
## # A tibble: 6 x 4
##   country      continent data              model
##   <fct>        <fct>     <list>            <list>
## 1 Algeria      Africa    <tibble [12 x 4]> <S3: lm>
## 2 Angola       Africa    <tibble [12 x 4]> <S3: lm>
## 3 Benin        Africa    <tibble [12 x 4]> <S3: lm>
## 4 Botswana     Africa    <tibble [12 x 4]> <S3: lm>
## 5 Burkina Faso Africa    <tibble [12 x 4]> <S3: lm>
## 6 Burundi      Africa    <tibble [12 x 4]> <S3: lm>
```

Transform + Visualise + Model

```
by_country <- by_country %>%
  mutate(
    resids = map2(data, model, add_residuals)
  )

head(by_country)
```

```
## # A tibble: 6 x 5
##   country     continent data              model    resids
##   <fct>       <fct>     <list>            <list>   <list>
## 1 Afghanistan Asia      <tibble [12 x 4]> <S3: lm> <tibble [12 x 5]>
## 2 Albania     Europe    <tibble [12 x 4]> <S3: lm> <tibble [12 x 5]>
## 3 Algeria     Africa    <tibble [12 x 4]> <S3: lm> <tibble [12 x 5]>
## 4 Angola      Africa    <tibble [12 x 4]> <S3: lm> <tibble [12 x 5]>
## 5 Argentina   Americas  <tibble [12 x 4]> <S3: lm> <tibble [12 x 5]>
## 6 Australia   Oceania   <tibble [12 x 4]> <S3: lm> <tibble [12 x 5]>
```

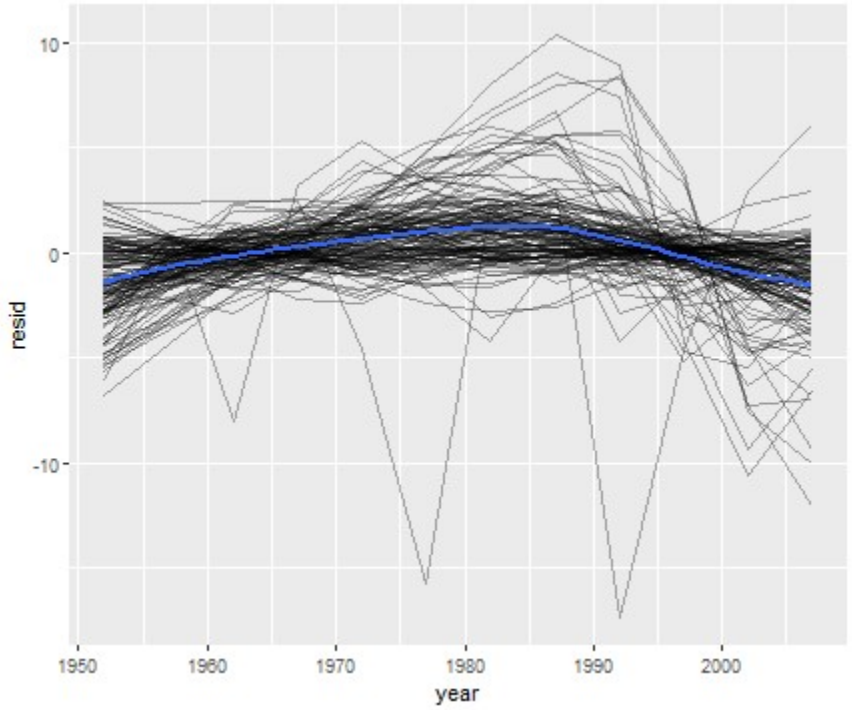
Model Residuals

```
resids <- unnest(by_country, resids)

resids
```

```
## # A tibble: 1,704 x 7
##    country     continent  year lifeExp      pop gdpPercap   resid
##    <fct>       <fct>     <int>   <dbl>    <int>     <dbl>   <dbl>
##  1 Afghanistan Asia       1952    28.8  8425333      779. -1.11
##  2 Afghanistan Asia       1957    30.3  9240934      821. -0.952
##  3 Afghanistan Asia       1962    32.0 10267083      853. -0.664
##  4 Afghanistan Asia       1967    34.0 11537966      836. -0.0172
##  5 Afghanistan Asia       1972    36.1 13079460      740.  0.674
##  6 Afghanistan Asia       1977    38.4 14880372      786.  1.65
##  7 Afghanistan Asia       1982    39.9 12881816      978.  1.69
##  8 Afghanistan Asia       1987    40.8 13867957      852.  1.28
##  9 Afghanistan Asia       1992    41.7 16317921      649.  0.754
## 10 Afghanistan Asia       1997    41.8 22227415      635. -0.534
## # ... with 1,694 more rows
```

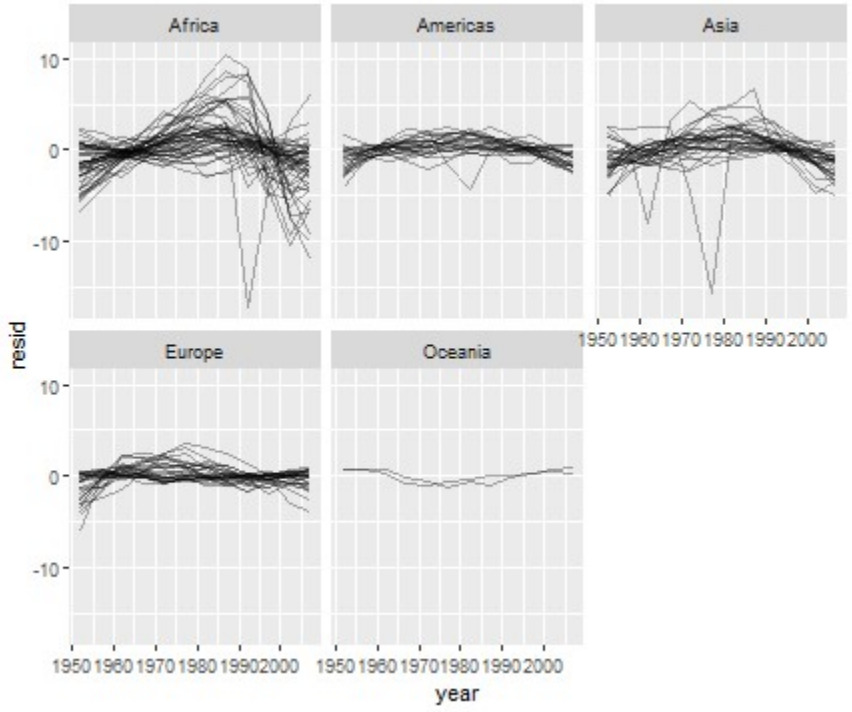
Transform + Visualise + Model

```
resids %>%
  ggplot(aes(year, resid)) +
    geom_line(aes(group = country),
                        alpha = 1 / 3) +
    geom_smooth(se = FALSE)
```



Transform + Visualise + Model

```
resids %>%
  ggplot(aes(year, resid, group = country)) +
    geom_line(alpha = 1 / 3) +
    facet_wrap(~continent)
```



Transform + Visualise + Model

```
by_country %>%
  mutate(glance = map(model, broom::glance)) %>%
  unnest(glance) %>%
  head()
```

```
## # A tibble: 6 x 16
##   country continent data  model resids r.squared adj.r.squared sigma
##   <fct>   <fct>     <lis> <lis> <list>     <dbl>         <dbl> <dbl>
## 1 Afghan~ Asia      <tib~ <S3:~ <tibb~     0.948         0.942 1.22
## 2 Albania Europe    <tib~ <S3:~ <tibb~     0.911         0.902 1.98
## 3 Algeria Africa    <tib~ <S3:~ <tibb~     0.985         0.984 1.32
## 4 Angola  Africa    <tib~ <S3:~ <tibb~     0.888         0.877 1.41
## 5 Argent~ Americas  <tib~ <S3:~ <tibb~     0.996         0.995 0.292
## 6 Austra~ Oceania   <tib~ <S3:~ <tibb~     0.980         0.978 0.621
## # ... with 8 more variables: statistic <dbl>, p.value <dbl>, df <int>,
## #   logLik <dbl>, AIC <dbl>, BIC <dbl>, deviance <dbl>, df.residual <int>
```

Transform + Visualise + Model

```
glance <- by_country %>%
  mutate(glance = map(model, broom::glance)) %>%
  unnest(glance, .drop = TRUE)
head(glance)
```

```
## # A tibble: 6 x 13
##   country continent r.squared adj.r.squared sigma statistic  p.value    df
##   <fct>   <fct>         <dbl>         <dbl> <dbl>     <dbl>    <dbl> <int>
## 1 Afghan~ Asia          0.948         0.942 1.22      181.  9.84e- 8     2
## 2 Albania Europe        0.911         0.902 1.98      102.  1.46e- 6     2
## 3 Algeria Africa        0.985         0.984 1.32      662.  1.81e-10     2
## 4 Angola  Africa        0.888         0.877 1.41       79.1 4.59e- 6     2
## 5 Argent~ Americas      0.996         0.995 0.292    2246.  4.22e-13     2
## 6 Austra~ Oceania       0.980         0.978 0.621     481.  8.67e-10     2
## # ... with 5 more variables: logLik <dbl>, AIC <dbl>, BIC <dbl>,
## #   deviance <dbl>, df.residual <int>
```

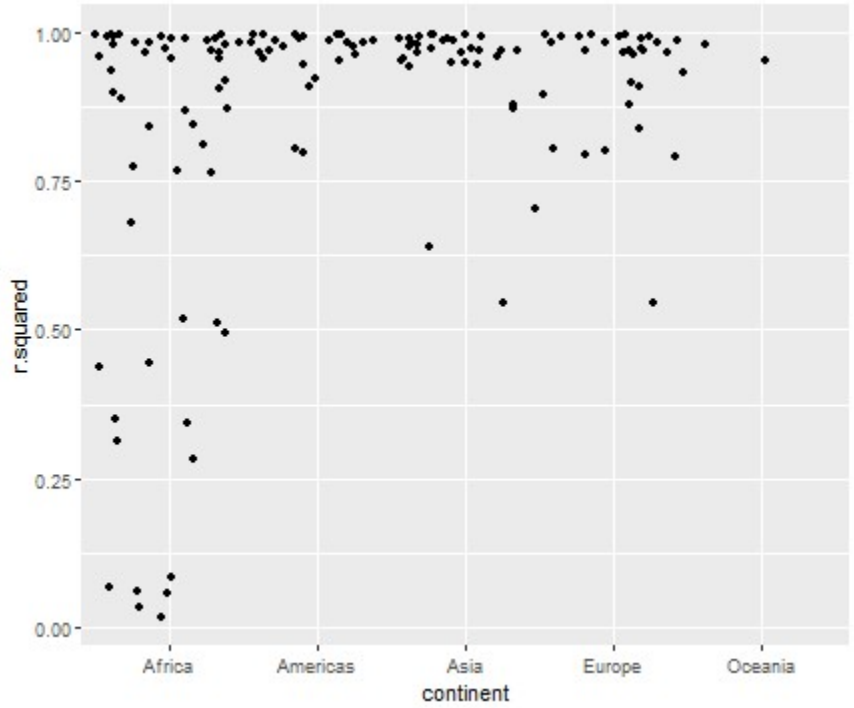
Model - Life Expectancy

```
glance %>%
  arrange(r.squared)
```

```
## # A tibble: 142 x 13
##    country continent r.squared adj.r.squared sigma statistic p.value    df
##    <fct>   <fct>         <dbl>         <dbl> <dbl>     <dbl>   <dbl> <int>
##  1 Rwanda  Africa       0.0172      -0.0811   6.56     0.175  0.685      2
##  2 Botswa~ Africa       0.0340      -0.0626   6.11     0.352  0.566      2
##  3 Zimbab~ Africa       0.0562      -0.0381   7.21     0.596  0.458      2
##  4 Zambia  Africa       0.0598      -0.0342   4.53     0.636  0.444      2
##  5 Swazil~ Africa       0.0682      -0.0250   6.64     0.732  0.412      2
##  6 Lesotho Africa       0.0849      -0.00666  5.93     0.927  0.358      2
##  7 Cote d~ Africa       0.283        0.212    3.93     3.95   0.0748     2
##  8 South ~ Africa       0.312        0.244    4.74     4.54   0.0588     2
##  9 Uganda  Africa       0.342        0.276    3.19     5.20   0.0457     2
## 10 Congo,~ Africa       0.348        0.283    2.43     5.34   0.0434     2
## # ... with 132 more rows, and 5 more variables: logLik <dbl>, AIC <dbl>,
## #   BIC <dbl>, deviance <dbl>, df.residual <int>
```

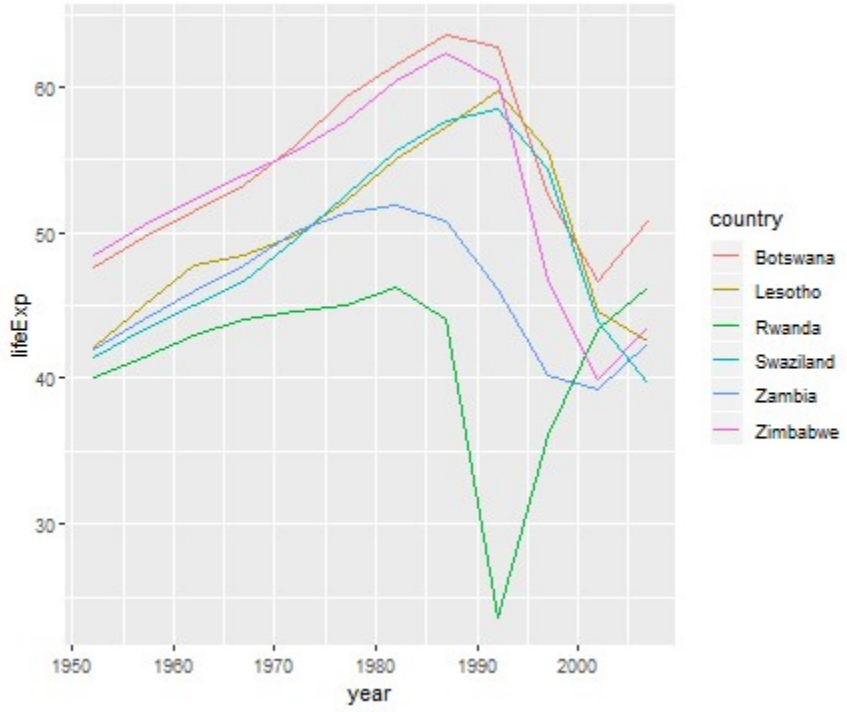
Transform + Visualise + Model

```
glance %>%
  ggplot(aes(continent, r.squared)) +
    geom_jitter(width = 0.5)
```



Transform + Visualise + Model

```
bad_fit <- filter(glance, r.squared < 0.25)

gapminder %>%
  semi_join(bad_fit, by = "country") %>%
  ggplot(aes(year, lifeExp, colour = countr
    geom_line()
```



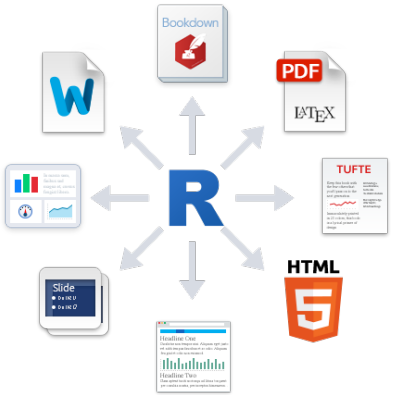
Communicate

Communication

- Reproducable documents, presentations, and reports
- Interactive web-apps and dashboards

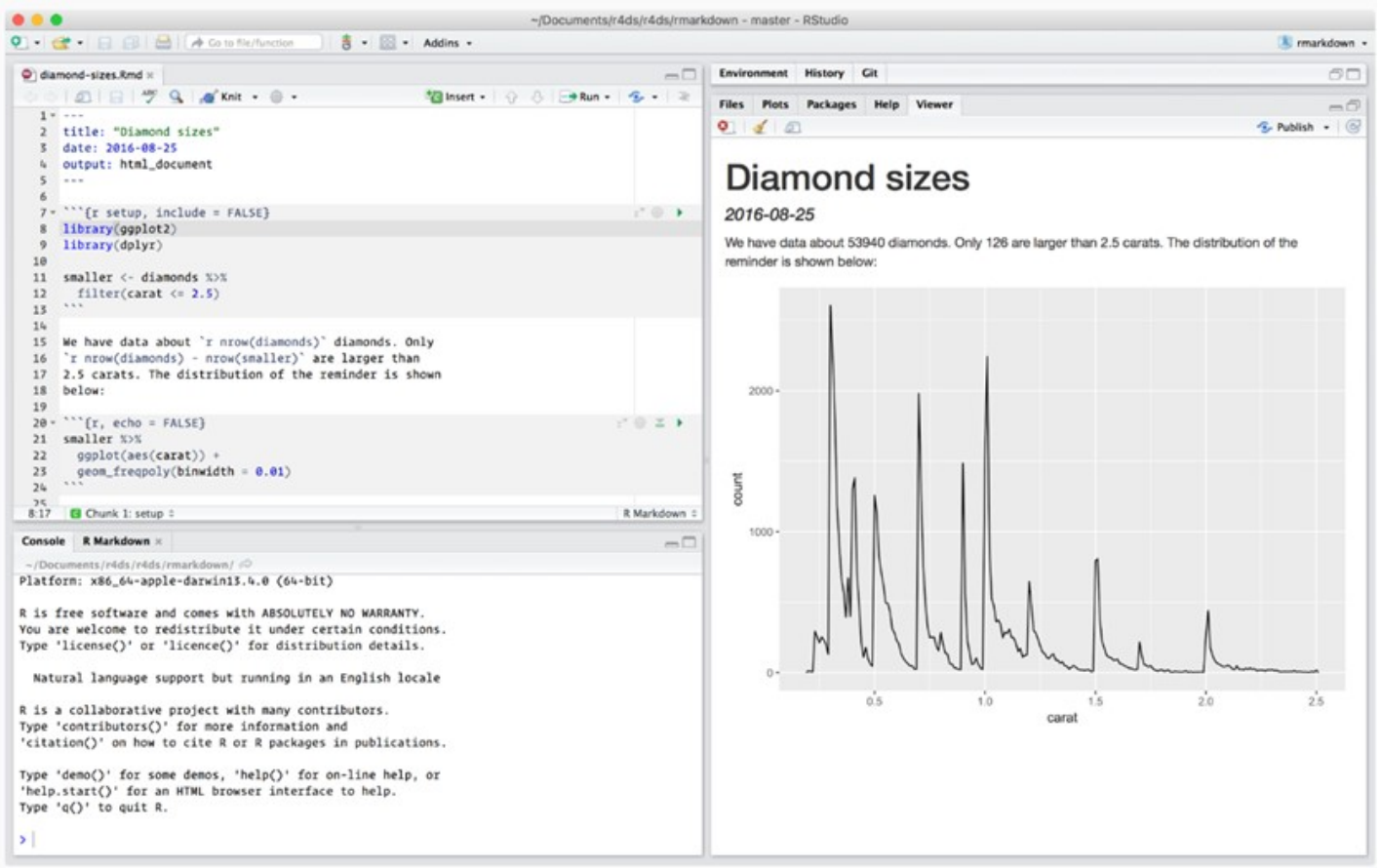
Rmarkdown

- Output Formats



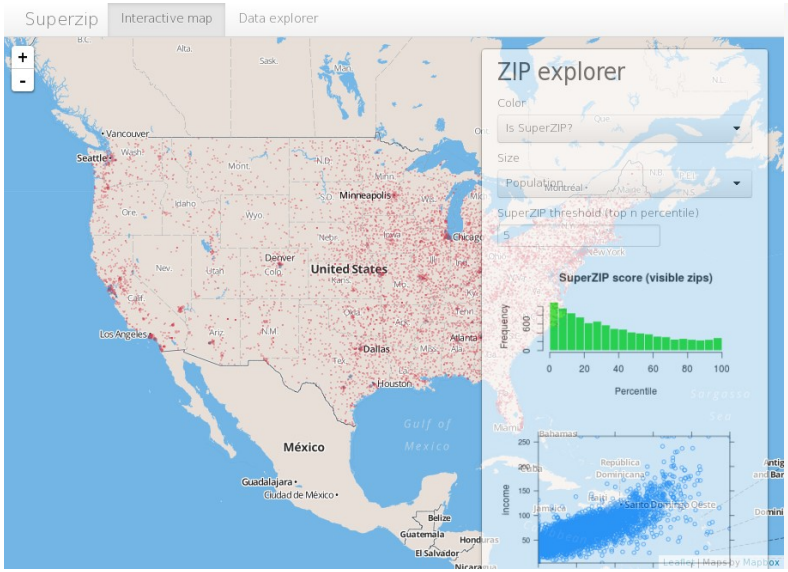
- A reproducible workflow

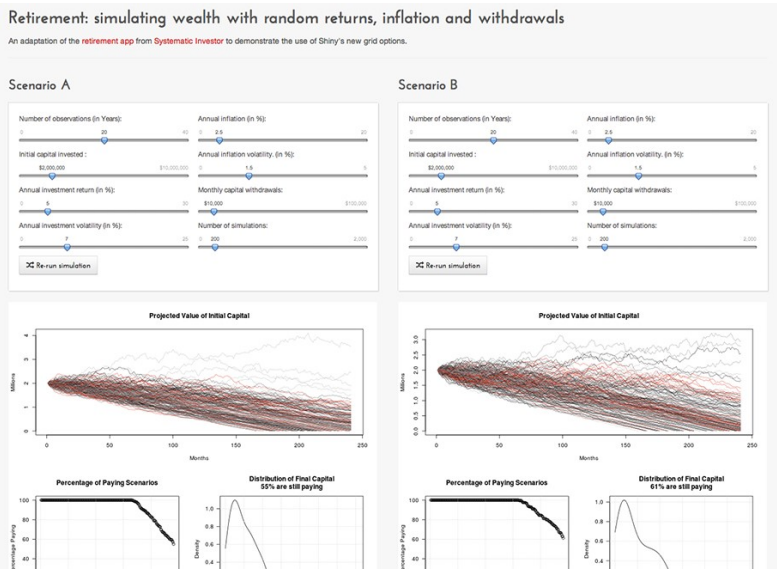
Example of rmarkdown report



Shiny

- Shiny combines the computational power of R with the interactivity of the modern web.





Shiny Dashboard

Questions ?

Resources 🎓

Resources

Not So Standard Deviations

Hilary Parker @hspter

Roger Peng @rdpeng

Credibly Curious

Saskia Freytag @trashystats

Nicholas Tierney @nj_tierney

DataFramed

Hugo Bowne-Anderson @hugobowne

Resources

Not So Standard Deviations

Hilary Parker @hspter

Roger Peng @rdpeng

Credibly Curious

Saskia Freytag @trashystats

Nicholas Tierney @nj_tierney

DataFramed

Hugo Bowne-Anderson @hugobowne

R for Data Science

Garrett Grolemund @StatGarrett

Hadley Wickham @hadleywickham

RStudio Community

A community for all things R and RStudio.

Stack Overflow

Community for developers to learn, share their knowledge, and build their careers.

R-bloggers

The site helps R bloggers and users to connect and follow the "R blogosphere"

Resources

Not So Standard Deviations

Hilary Parker @hspter

Roger Peng @rdpeng

Credibly Curious

Saskia Freytag @trashystats

Nicholas Tierney @nj_tierney

DataFramed

Hugo Bowne-Anderson @hugobowne

R for Data Science

Garrett Grolemund @StatGarrett

Hadley Wickham @hadleywickham

RStudio Community

A community for all things R and RStudio.

Stack Overflow

Community for developers to learn, share their knowledge, and build their careers.

R-bloggers

The site helps R bloggers and users to connect and follow the "R blogosphere"

RUsers

Johannesburg

Cape Town

R-Ladies

R-Ladies Community Slack Workspace

Johannesburg

Cape Town

Thank you! - End of File