

Introduction to R Workshop

Session 3
Sean Nguyen



MSU > BEST

Broadening Experiences in Scientific Training

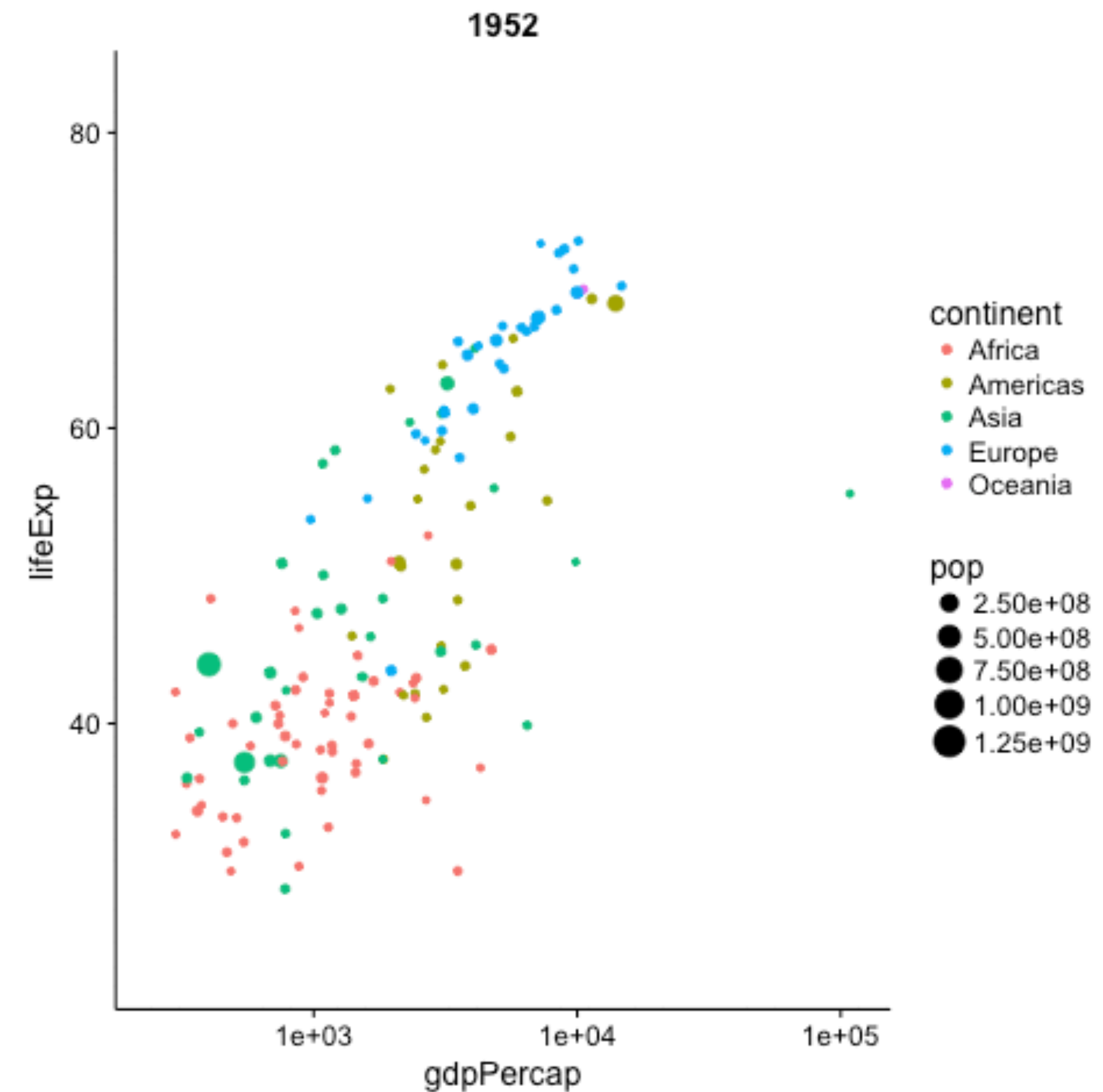
Session 3: Goals

Learn factors and levels

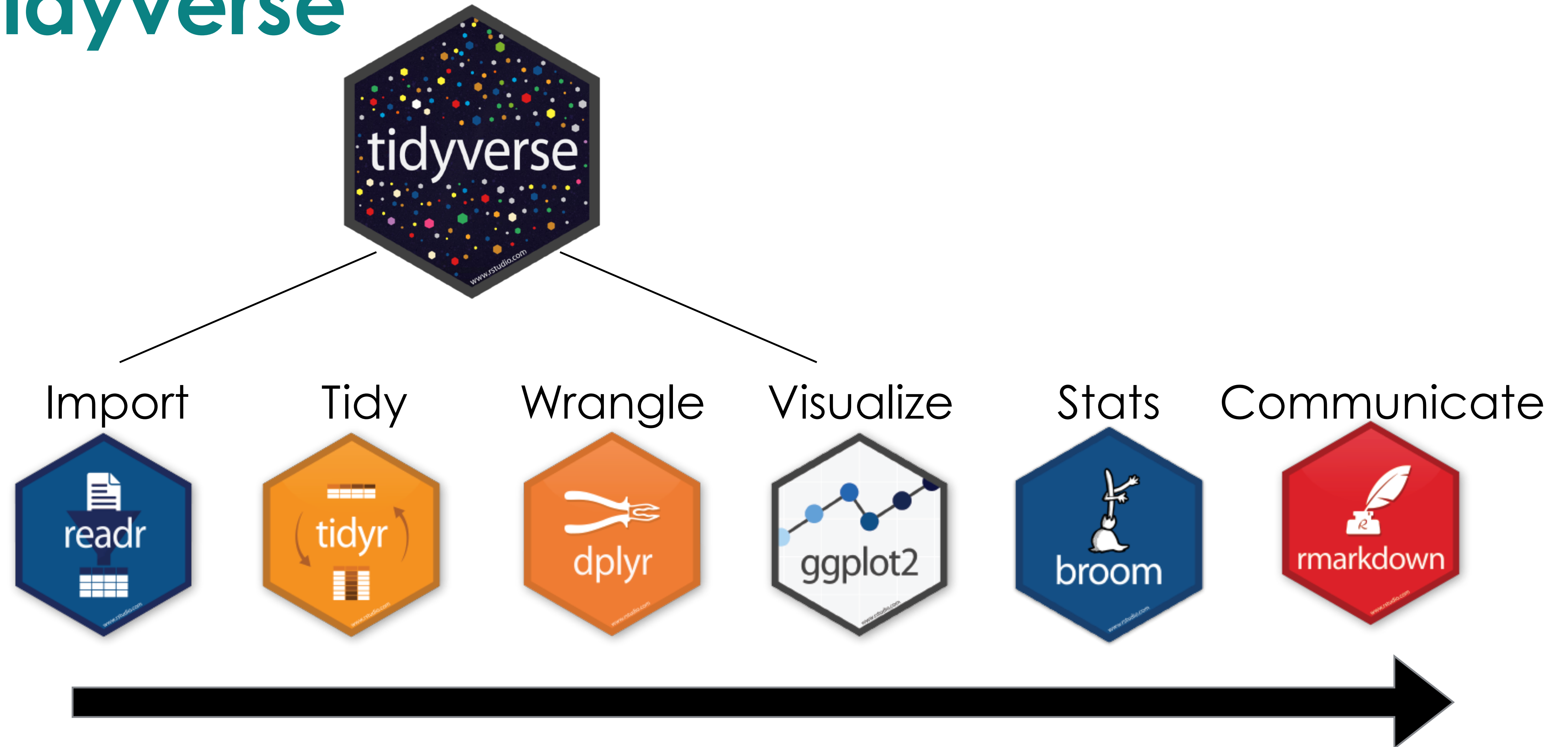
ggplot2

Saving plots

R markdown



Data Analysis in the Tidyverse



Data Analysis in the Tidyverse

Import



Tidy



Wrangle



Visualize



```
ggplot()  
geom_line()  
geom_bar()  
geom_point()  
geom_boxplot()
```

Stats



Communicate



```
knitr  
Rmd  
Markdown
```



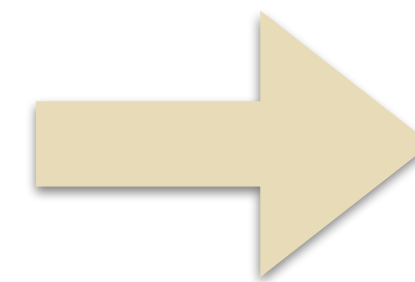
factor: categorical variable



Formula: `data$column <- as.factor(data$column)`

`data$month <- as.factor(data$month)`

month (chr)
January
February
March



month (fctr)
January
February
March

factor: categorical variable

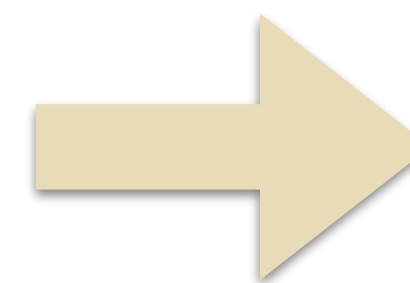


Formula for

multiple columns: `mutate_at(vars(columnA:columnD), as.factor)`

`mutate_at(vars(month:group), as.factor)`

month (chr)	day (chr)	group (chr)
January	Mon	A
February	Mon	A
March	Tues	B



month (fctr)	day (fctr)	group (fctr)
January	Mon	A
February	Mon	A
March	Tues	B

numeric: numeric variable



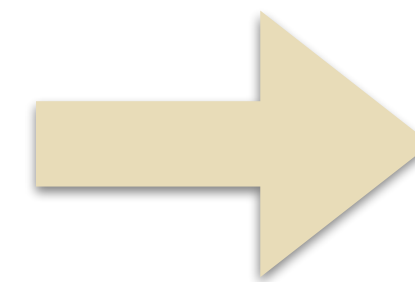
Integers - (~2 billion)

Double - (~1.79e308)

1.15, 4.40, 3.80

Formula: `data$column <- as.numeric(data$column)`

Column (chr)
3.124
5.934
5.600



Column (dbl)
3.124
5.934
5.600

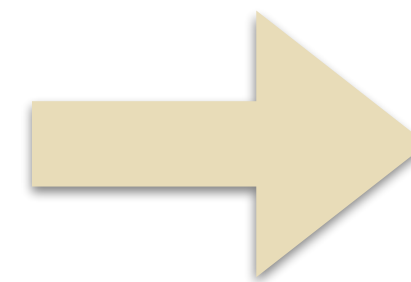
character: strings of text



“MSU R Workshop”

Formula: `data$column <- as.character(data$column)`

Name (fctr)
Debbie
Dylan
Sarah



Name (chr)
Debbie
Dylan
Sarah

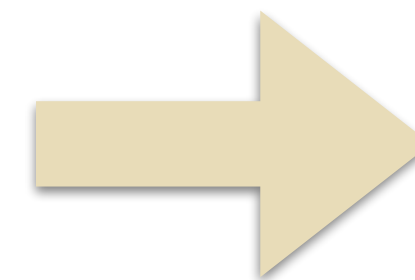
logical: True/False variable



Formula: `data$column <- as.logical(data$column)`

`data$question <- as.logical(data$question)`

question (chr)
TRUE
FALSE
TRUE



question (lgl)
TRUE
FALSE
TRUE

data types in R:

Factors - categorical variable Monday , Tuesday, Wednesday

Numeric - numbers

Integers (~2 billion)

Double (~1.79e308)

1.15, 4.40, 3.80



Character - strings of characters

“Michigan State University R Workshop”

Logical - TRUE/FALSE

ggplot2 - considerations

tidy data - variables in columns, observations in rows

factors - categorical variables

integers/numeric - number variables

levels - order of categorical variables



country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174608398
China	1999	212258	127291272
China	2000	216766	128042583

variables

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174608398
China	1999	212258	127291272
China	2000	216766	128042583

observations

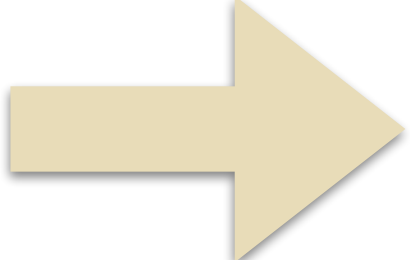
country	year	cases	population
Afghanistan	99	745	19987071
Afghanistan	00	2666	20595360
Brazil	99	37737	172006362
Brazil	00	80488	174608398
China	99	212258	127291272
China	00	216766	128042583

values

levels - the order of categorical variables

Formula: `data$column <- factor(data$column,
levels = c("order", "that", "you", "want"))`

`data$month <- factor(data$month, levels = c("January", "February",
"March", "April"))`

`'April', 'February', 'January', 'March'`  `'January', 'February', 'March', 'April'`

ggplot2 - powerful plotting library

General Formula: `ggplot(aes(x = ____, y = ____)) +`

`geom_point()`

`geom_line()`

`geom_boxplot()`

`geom_violin()`

`geom_col()/geom_bar(stat = "identity")`

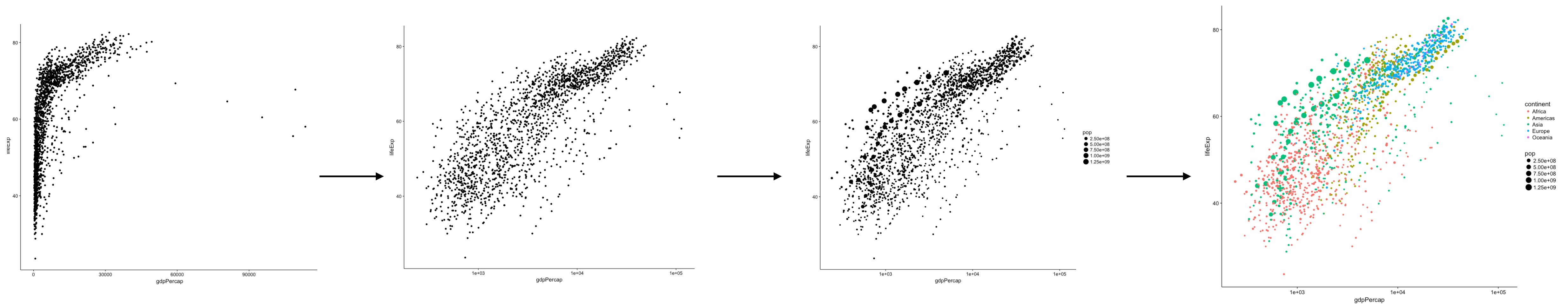


Grammar of Graphics



- Layered
- Iterative
- Customizable

```
# A tibble: 1,704 × 6
  country continent year lifeExp   pop gdpPercap
  <fctr>   <fctr> <int>   <dbl> <int>   <dbl>
1 Afghanistan Asia  1952  28.801 8425333  779.4453
2 Afghanistan Asia  1957  30.332 9240934  820.8530
3 Afghanistan Asia  1962  31.997 10267083  853.1007
4 Afghanistan Asia  1967  34.020 11537966  836.1971
5 Afghanistan Asia  1972  36.088 13079460  739.9811
```



```
data %>%
  ggplot(aes(x = gdpPercap, y = lifeExp))+
  geom_point()
```

```
data %>%
  ggplot(aes(x = gdpPercap, y = lifeExp))+
  geom_point()+
  scale_x_log10()
```

```
data %>%
  ggplot(aes(x = gdpPercap, y = lifeExp, size = pop))+
  geom_point()+
  scale_x_log10()
```

```
data %>%
  ggplot(aes(x = gdpPercap, y = lifeExp,
            size = pop, color = continent))+
  geom_point()+
  scale_x_log10()
```


Demo!

Try to plot:

Life expectancy of Asian countries in 1992

Life expectancy of of Africa and Europe in 2007

GDP of Americas and Europe in 2002

BONUS:

Determine the GDP if each continent in 2007

ggplot2 - tweaking your plot



reorder - `ggplot(aes (x = reorder(_____ , ordered_variable), y = _____)`

color - `+ scale_fill_brewer(palette= "YlOrRd")`

legends - `remove legend + guides(color = FALSE)/guides(fill = FALSE)`

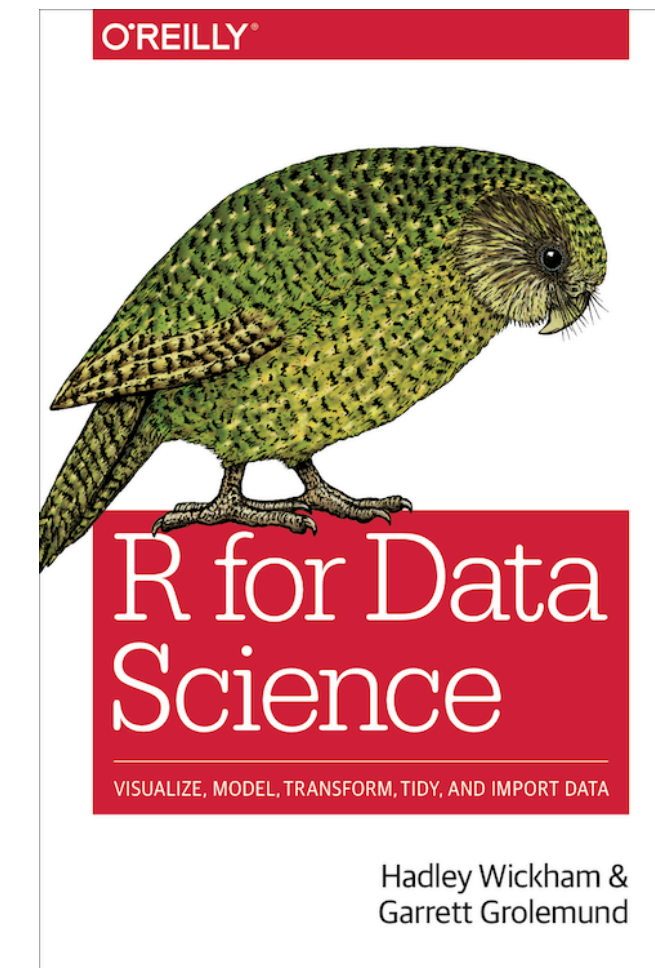
annotations - `+ annotate("text",x= 2, y=2, label = "your text")`

Additional Resources

R Graphics Cookbook



R for Data Science



Stack Overflow



stackoverflow

R markdown: reproducible documents

Analyze.

Share.

Reproduce.

