# Exploring Data with R

Tony Yao-Jen Kuo

# Overview

# Introducing the tidyverse system

- Picked by RStudio
- `dplyr` for data manipulation
- `ggplot` for data visualization
- And more. . .

# We are gonna talk about 3 packages

- `gapminder` for data
- `dplyr`
- `ggplot2`

# gapminder

# Getting data

```
file_url <- "https://storage.googleapis.com/learn_pd_like_t
gap_minder <- read.csv(file_url, stringsAsFactors = FALSE)
```

# The story of Hans Rosling and Gapminder

https://youtu.be/jbkSRLYSojo

# dplyr

# Installing dplyr

```r
install.packages("dplyr")
```

# Basic functions

- `filter()`

# Basic functions

- `filter()`
- `select()`

# Basic functions

- `filter()`
- `select()`
- `arrange()`

# Basic functions

- `filter()`
- `select()`
- `arrange()`
- `mutate()`

# Basic functions

- `filter()`
- `select()`
- `arrange()`
- `mutate()`
- `summarise()`

# Basic functions

- `filter()`
- `select()`
- `arrange()`
- `mutate()`
- `summarise()`
- `group_by()`

# filter() for subsetting rows

```r
library(dplyr)

gap_minder %>%
  filter(country == "Taiwan")
```

```
##    country continent year lifeExp      pop gdpPercap
## 1   Taiwan      Asia 1952   58.50  8550362  1206.948
## 2   Taiwan      Asia 1957   62.40 10164215  1507.861
## 3   Taiwan      Asia 1962   65.20 11918938  1822.879
## 4   Taiwan      Asia 1967   67.50 13648692  2643.859
## 5   Taiwan      Asia 1972   69.39 15226039  4062.524
## 6   Taiwan      Asia 1977   70.59 16785196  5596.520
## 7   Taiwan      Asia 1982   72.16 18501390  7426.355
## 8   Taiwan      Asia 1987   73.40 19757799 11054.562
## 9   Taiwan      Asia 1992   74.26 20686918 15215.658
## 10  Taiwan      Asia 1997   75.25 21628605 20206.821
## 11  Taiwan      Asia 2002   76.99 22454239 23235.423
## 12  Taiwan      Asia 2007   78.40 23174294 28718.277
```

# select() for extracting columns

```
gap_minder %>%
  filter(country == "Taiwan") %>%
  select(year, gdpPercap, lifeExp)
```

```
##    year gdpPercap lifeExp
## 1  1952  1206.948   58.50
## 2  1957  1507.861   62.40
## 3  1962  1822.879   65.20
## 4  1967  2643.859   67.50
## 5  1972  4062.524   69.39
## 6  1977  5596.520   70.59
## 7  1982  7426.355   72.16
## 8  1987 11054.562   73.40
## 9  1992 15215.658   74.26
## 10 1997 20206.821   75.25
## 11 2002 23235.423   76.99
## 12 2007 28718.277   78.40
```

# arrange() for sorting

```
gap_minder %>%
  filter(continent == "Asia") %>%
  filter(year == 2007) %>%
  arrange(gdpPercap)
```

```
##                  country continent year lifeExp        pop
## 1                Myanmar      Asia 2007  62.069   47761980
## 2            Afghanistan      Asia 2007  43.828   31889923
## 3                  Nepal      Asia 2007  63.785   28901790
## 4             Bangladesh      Asia 2007  64.062  150448339
## 5       Korea, Dem. Rep.      Asia 2007  67.297   23301725
## 6               Cambodia      Asia 2007  59.723   14131858
## 7            Yemen, Rep.      Asia 2007  62.698   22211743
## 8                Vietnam      Asia 2007  74.249   85262356
## 9                  India      Asia 2007  64.698 1110396331
## 10              Pakistan      Asia 2007  65.483  169270617
## 11    West Bank and Gaza      Asia 2007  73.422    4018332
## 12              Mongolia      Asia 2007  66.803    2874127
```

# mutate() for creating new columns

```r
gap_minder %>%
  filter(country == "Taiwan") %>%
  mutate(gdp_million = (gdpPercap * pop / 1000000))
```

```
##    country continent year lifeExp      pop gdpPercap gdp
## 1   Taiwan      Asia 1952   58.50  8550362  1206.948
## 2   Taiwan      Asia 1957   62.40 10164215  1507.861
## 3   Taiwan      Asia 1962   65.20 11918938  1822.879
## 4   Taiwan      Asia 1967   67.50 13648692  2643.859
## 5   Taiwan      Asia 1972   69.39 15226039  4062.524
## 6   Taiwan      Asia 1977   70.59 16785196  5596.520
## 7   Taiwan      Asia 1982   72.16 18501390  7426.355   1
## 8   Taiwan      Asia 1987   73.40 19757799 11054.562   2
## 9   Taiwan      Asia 1992   74.26 20686918 15215.658   3
## 10  Taiwan      Asia 1997   75.25 21628605 20206.821   4
## 11  Taiwan      Asia 2002   76.99 22454239 23235.423   5
## 12  Taiwan      Asia 2007   78.40 23174294 28718.277   6
```

# summarise() for…a summary

```
gap_minder %>%
  summarise(median(gdpPercap))
```

```
##   median(gdpPercap)
## 1          3531.847
```

# group_by() for a grouped summary

```
gap_minder %>%
  group_by(continent) %>%
  summarise(medianGdpPercap = median(gdpPercap))
```

```
## # A tibble: 5 x 2
##   continent medianGdpPercap
##   <chr>               <dbl>
## 1 Africa              1192.
## 2 Americas            5466.
## 3 Asia                2647.
## 4 Europe             12082.
## 5 Oceania            17983.
```

# Going further

https://dplyr.tidyverse.org/

# ggplot2

# gg stands for. . .

*The grammar of graphics.*

# Installing ggplot2

```r
install.packages("ggplot2")
```

# Basic concepts

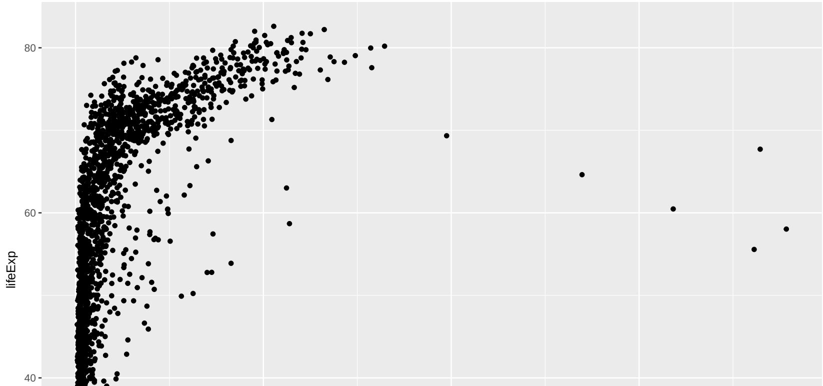
- `ggplot(aes(x = , y = , color = , fill = , ...))` for data mapping
- `geom_OOO()` for different charts‘
- Using + to add different layers

# geom_point() for exploring correlations

Making a scatter plot

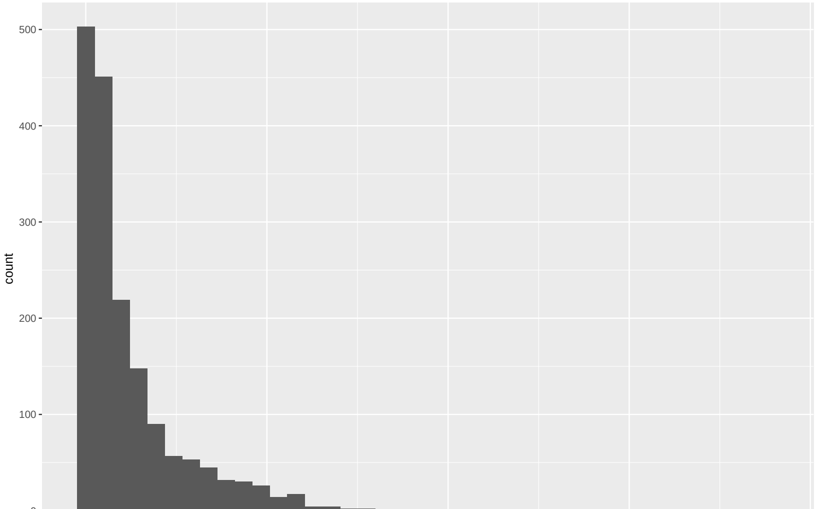
```
library(ggplot2)

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



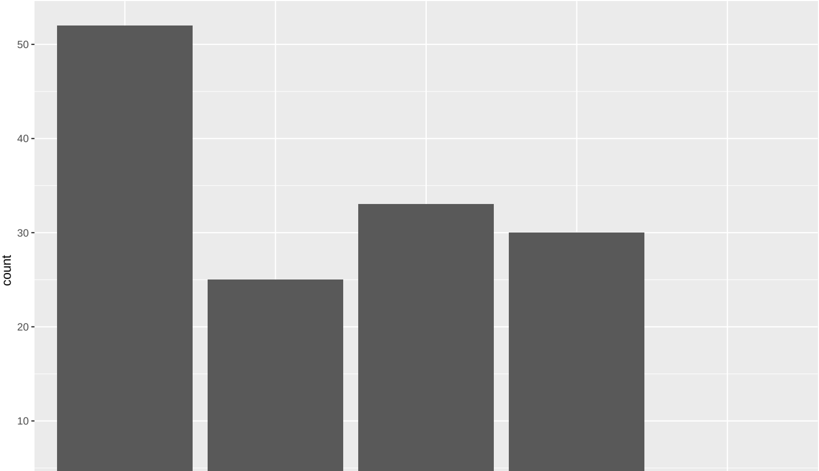
# geom_histogram() for exploring distributions

```
gap_minder %>%
  ggplot(aes(x = gdpPercap)) +
  geom_histogram(bins = 40)
```



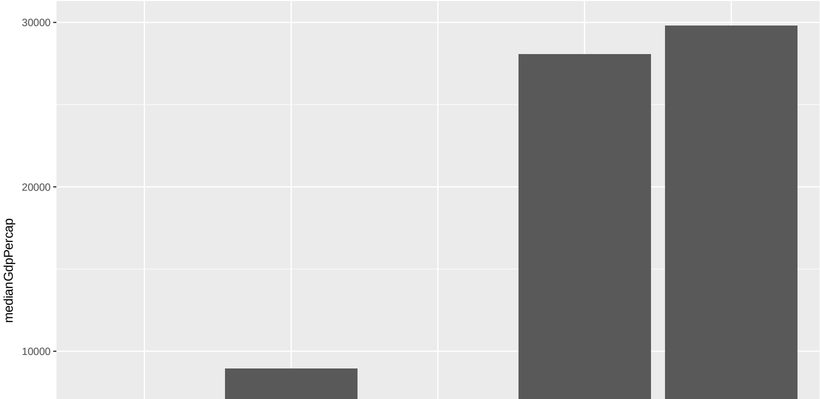
# geom_bar() for exploring row counts

```
gap_minder %>%
  filter(year == 2007) %>%
  ggplot(aes(x = continent)) +
  geom_bar()
```



# geom_bar() for grouped summary

```
gap_minder %>%
  filter(year == 2007) %>%
  group_by(continent) %>%
  summarise(medianGdpPercap = median(gdpPercap)) %>%
  ggplot(aes(x = continent, y = medianGdpPercap)) +
  geom_bar(stat = "identity")
```



# Going further

https://ggplot2.tidyverse.org/

# Animated plot for inspirations

# Installing plotly

```r
install.packages("plotly")
```

# Plotting a gapminder replica

```r
library(plotly)
radius <- sqrt((gap_minder$pop)/pi)

p <- gap_minder %>%
  plot_ly(
    x = ~gdpPercap,
    y = ~lifeExp,
    size = ~pop,
    color = ~continent,
    frame = ~year,
    text = ~country,
    hoverinfo = "text",
    type = 'scatter',
    mode = 'markers',
    sizes = c(min(radius), max(radius))
  ) %>%
  layout(
    xaxis = list(
```

# The gapminder replica

```
p
```