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 a data gapminder
- `dplyr` for grammar of data manipulation
- `ggplot2` for grammar of graphics

gapminder

Getting data

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

```r
library(gapminder)
```

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)

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

```
## # A tibble: 12 x 6
##    country continent  year lifeExp      pop gdpPercap
##    <fct>   <fct>     <int>   <dbl>    <int>     <dbl>
##  1 Taiwan  Asia       1952    58.5  8550362     1207.
##  2 Taiwan  Asia       1957    62.4 10164215     1508.
##  3 Taiwan  Asia       1962    65.2 11918938     1823.
##  4 Taiwan  Asia       1967    67.5 13648692     2644.
##  5 Taiwan  Asia       1972    69.4 15226039     4063.
##  6 Taiwan  Asia       1977    70.6 16785196     5597.
##  7 Taiwan  Asia       1982    72.2 18501390     7426.
##  8 Taiwan  Asia       1987    73.4 19757799    11055.
##  9 Taiwan  Asia       1992    74.3 20686918    15216.
## 10 Taiwan  Asia       1997    75.2 21628605    20207.
```

select() for extracting columns

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

```
## # A tibble: 12 x 3
##     year gdpPercap lifeExp
##    <int>     <dbl>   <dbl>
##  1  1952     1207.    58.5
##  2  1957     1508.    62.4
##  3  1962     1823.    65.2
##  4  1967     2644.    67.5
##  5  1972     4063.    69.4
##  6  1977     5597.    70.6
##  7  1982     7426.    72.2
##  8  1987    11055.    73.4
##  9  1992    15216.    74.3
## 10  1997    20207.    75.2
## 11  2002    23235.    77.0
```

arrange() for sorting

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

```
## # A tibble: 33 x 6
##    country          continent  year lifeExp        pop g
##    <fct>            <fct>     <int>   <dbl>      <int>
##  1 Myanmar          Asia       2007    62.1   47761980
##  2 Afghanistan      Asia       2007    43.8   31889923
##  3 Nepal            Asia       2007    63.8   28901790
##  4 Bangladesh       Asia       2007    64.1  150448339
##  5 Korea, Dem. Rep. Asia       2007    67.3   23301725
##  6 Cambodia         Asia       2007    59.7   14131858
##  7 Yemen, Rep.      Asia       2007    62.7   22211743
##  8 Vietnam          Asia       2007    74.2   85262356
##  9 India            Asia       2007    64.7 1110396331
## 10 Pakistan         Asia       2007    65.5  169270617
```

mutate() for creating new columns

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

```
## # A tibble: 12 x 7
##    country continent  year lifeExp      pop gdpPercap gd
##    <fct>   <fct>     <int>   <dbl>    <int>     <dbl>
##  1 Taiwan  Asia       1952    58.5  8550362     1207.
##  2 Taiwan  Asia       1957    62.4 10164215     1508.
##  3 Taiwan  Asia       1962    65.2 11918938     1823.
##  4 Taiwan  Asia       1967    67.5 13648692     2644.
##  5 Taiwan  Asia       1972    69.4 15226039     4063.
##  6 Taiwan  Asia       1977    70.6 16785196     5597.
##  7 Taiwan  Asia       1982    72.2 18501390     7426.
##  8 Taiwan  Asia       1987    73.4 19757799    11055.
##  9 Taiwan  Asia       1992    74.3 20686918    15216.
## 10 Taiwan  Asia       1997    75.2 21628605    20207.
## 11 Taiwan  Asia       2002    77.0 22454239    23235.
```

summarise() for...a summary

```r
gapminder %>%
  summarise(median(gdpPercap))
```

```
## # A tibble: 1 x 1
##   `median(gdpPercap)`
##                 <dbl>
## 1               3532.
```

group_by() for a grouped summary

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

```
## # A tibble: 5 x 2
##   continent medianGdpPercap
##   <fct>               <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

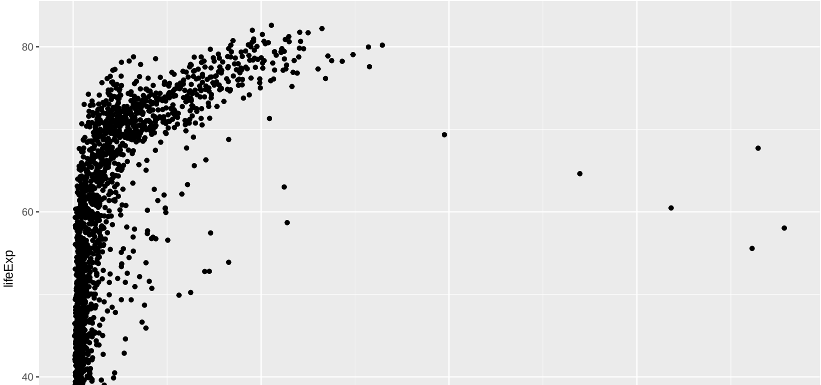
- `ggplot(aes(x = , y = , color = , fill = , ...))` for data mapping
- `geom_OOO()` for different charts‘
- Using + to add different layers

geom_point() for exploring correlations

Making a scatter plot

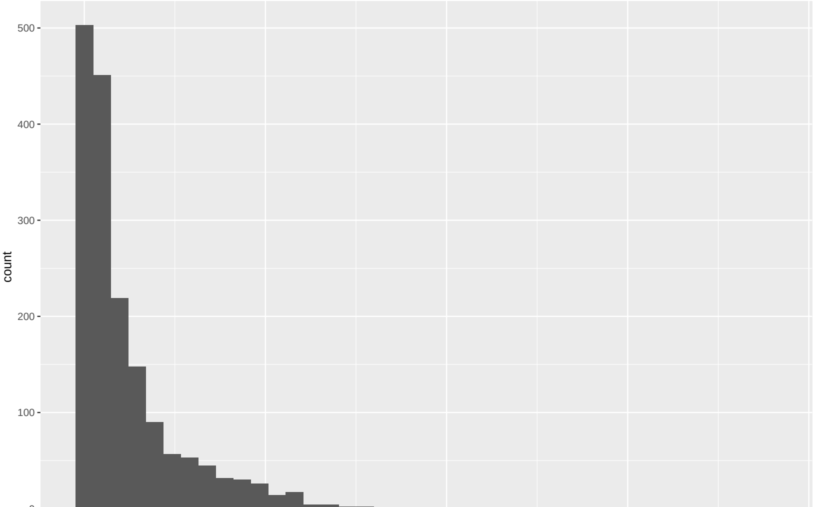
```
library(ggplot2)

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



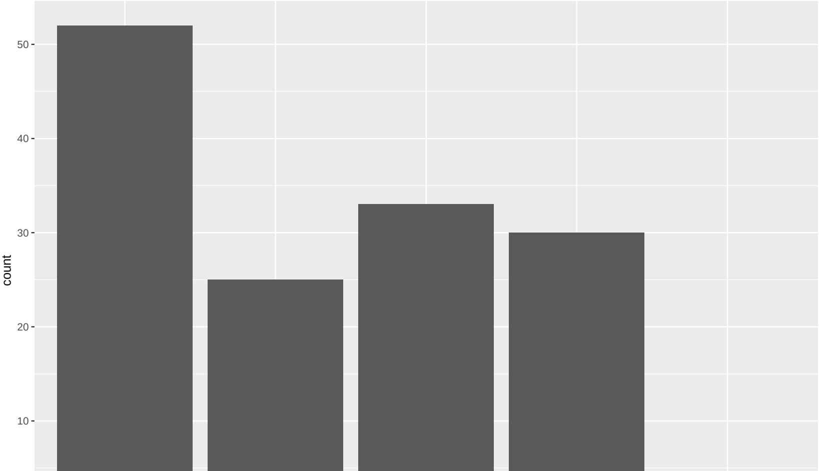
geom_histogram() for exploring distributions

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



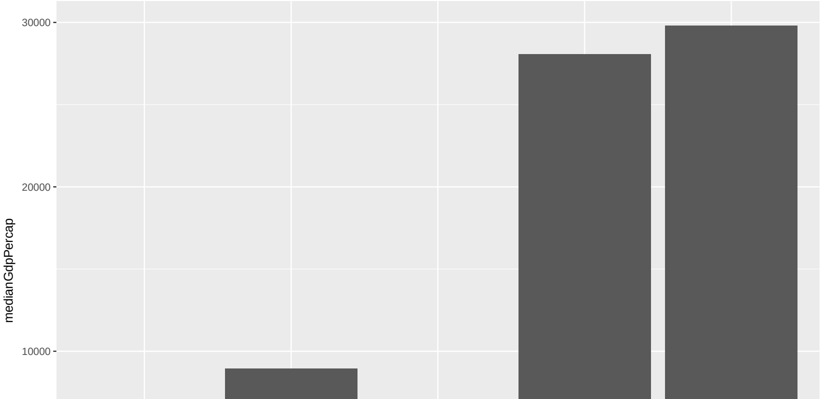
geom_bar() for exploring row counts

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



geom_bar() for grouped summary

```
gapminder %>%
  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

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

Plotting a gapminder replica

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

p <- gapminder %>%
  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
```