Section 4 Overview

In Section 4, you will look at a case study involving data from the Gapminder Foundation about trends in world health and economics.

After completing Section 4, you will:

- understand how Hans Rosling and the Gapminder Foundation use effective data visualization to convey data-based trends.
- be able to apply the **ggplot2** techniques from the previous section to answer questions using data.
- understand how fixed scales across plots can ease comparisons.
- be able to modify graphs to improve data visualization.

There is 1 assignment that uses the DataCamp platform for you to practice your coding skills.

We encourage you to use R to interactively test out your answers and further your learning.

Case Study: Trends in World Health and Economics

Case Study: Trends in World Health and Economics

"New Insights on Poverty" and "The Best Stats You've Ever Seen"

are the title of these talks.

Specifically, in this section, we set out

to answer the following two questions.

First, is it a fair characterization of today's world

to say that it is divided into a Western rich nations,

and the developing world in Africa, Asia, and Latin America?

Second, has income inequality across countries

worsened during the last 40 years?

We're going to use data

Video

Download video file

Transcripts

Download SubRip (.srt) file

Download Text (.txt) file

Textbook link

This video corresponds to the textbook section introducing the case study on new insights in poverty.

More about Gapminder

The original Gapminder TED talks are available and we encourage you to watch them.

- The Best Stats You've Ever Seen
- New Insights on Poverty

You can also find more information and raw data (in addition to what we analyze in class) at https://www.gapminder.org/.

Key points

- Data visualization can be used to dispel common myths and educate the public and contradict sensationalist or outdated claims and stories.
- We will use real data to answer the following questions about world health and economics:
    - Is it still fair to consider the world as divided into the West and the developing world?
    - Has income inequality across countries worsened over the last 40 years?

Learn About Verified Certificates

Gapminder Dataset

Gapminder Dataset

0:00 / 0:00 1.0x

knowledge regarding differences in child

mortality across different countries.

To get us started, we're going to take a quiz created

by Hans Rosling in his video New Insights on Poverty,

and we're going to start by testing our knowledge regarding differences

in child mortality across different countries.

So here's a quiz.

For each of the pairs of countries here, which country

do you think had the highest child mortality in 2015?

Video

Download video file

Transcripts

Download SubRip (.srt) file

Download Text (.txt) file

Textbook link

This video corresponds to the textbook section introducing the case study on new insights on poverty.

Key points

- A selection of world health and economics statistics from the Gapminder project can be found in the **dslabs** package as `data(gapminder)`.
- Most people have misconceptions about world health and economics, which can be addressed by considering real data.

Code

```
# load and inspect gapminder data
library(dslabs)
data(gapminder)
head(gapminder)

# compare infant mortality in Sri Lanka and Turkey
gapminder %>%
    filter(year == 2015 & country %in% c("Sri Lanka", "Turkey")) %>%
    select(country, infant_mortality)
```

Learn About Verified Certificates

Life Expectancy and Fertility Rates

Life Expectancy and Fertility Rates

Start of transcript. Skip to the end.

0:00 / 0:00 1.0x

RAFAEL IRIZARRY: Our misconceptions stem from the preconceived notion

that the world is divided into two groups, the Western World, composed

of Western Europe and North America, which

is characterized by long lifespans and small

Video

Download video file

Transcripts

Download SubRip (.srt) file

Download Text (.txt) file

Textbook link

This video corresponds to the textbook section on Gapminder scatterplots.

Key points

- A prevalent worldview is that the world is divided into two groups of countries:
    - Western world: high life expectancy, low fertility rate
    - Developing world: lower life expectancy, higher fertility rate
- Gapminder data can be used to evaluate the validity of this view.
- A scatterplot of life expectancy versus fertility rate in 1962 suggests that this viewpoint was grounded in reality 50 years ago. Is it still the case today?

Code

```
# basic scatterplot of life expectancy versus fertility
ds_theme_set()    # set plot theme
filter(gapminder, year == 1962) %>%
    ggplot(aes(fertility, life_expectancy)) +
    geom_point()

# add color as continent
filter(gapminder, year == 1962) %>%
    ggplot(aes(fertility, life_expectancy, color = continent)) +
    geom_point()
```

Learn About Verified Certificates

Faceting

Faceting

Start of transcript. Skip to the end.

0:00 / 0:00 1.0x

RAFAEL IRIZARRY: We could easily plot the 2012

data in the same way we did for 1962.

But for comparison, side by side plots are preferable.

In ggplot, we can achieve this by faceting variables.

We stratify the data by some variable and make the same plot for each strata

Video

Download video file

Transcripts

Download SubRip (.srt) file

Download Text (.txt) file

Textbook link

This video corresponds to the textbook section on faceting.

Key points

- Faceting makes multiple side-by-side plots stratified by some variable. This is a way to ease comparisons.
- The `facet_grid()` function allows faceting by up to two variables, with rows faceted by one variable and columns faceted by the other variable. To facet by only one variable, use the dot operator as the other variable.
- The `facet_wrap()` function facets by one variable and automatically wraps the series of plots so they have readable dimensions.
- Faceting keeps the axes fixed across all plots, easing comparisons between plots.
- The data suggest that the developing versus Western world view no longer makes sense in 2012.

Code

```
# facet by continent and year
filter(gapminder, year %in% c(1962, 2012)) %>%
    ggplot(aes(fertility, life_expectancy, col = continent)) +
    geom_point() +
    facet_grid(continent ~ year)

# facet by year only
filter(gapminder, year %in% c(1962, 2012)) %>%
    ggplot(aes(fertility, life_expectancy, col = continent)) +
    geom_point() +
    facet_grid(. ~ year)

# facet by year, plots wrapped onto multiple rows
years <- c(1962, 1980, 1990, 2000, 2012)
continents <- c("Europe", "Asia")
gapminder %>%
    filter(year %in% years & continent %in% continents) %>%
    ggplot(aes(fertility, life_expectancy, col = continent)) +
    geom_point() +
    facet_wrap(~year)
```

Learn About Verified Certificates

Time Series Plots

Time Series Plots

Start of transcript. Skip to the end.

0:00 / 0:00 1.0x

RAFAEL IRIZARRY: The visualizations we have just seen

effectively illustrate that data no longer

supports the Western versus developing worldview.

But once we see these plots, new questions emerge.

For example, which countries are improving

Video

Download video file

Transcripts

Download SubRip (.srt) file

Download Text (.txt) file

Textbook link

This video corresponds to the textbook section on time series plots.

Key points

- Time series plots have time on the x-axis and a variable of interest on the y-axis.
- The `geom_line()` geometry connects adjacent data points to form a continuous line. A line plot is appropriate when points are regularly spaced, densely packed and from a single data series.
- You can plot multiple lines on the same graph. Remember to group or color by a variable so that the lines are plotted independently.
- Labeling is usually preferred over legends. However, legends are easier to make and appear by default. Add a label with `geom_text()`, specifying the coordinates where the label should appear on the graph.

Code: Single time series

```
# scatterplot of US fertility by year
gapminder %>%
    filter(country == "United States") %>%
    ggplot(aes(year, fertility)) +
    geom_point()

# line plot of US fertility by year
gapminder %>%
    filter(country == "United States") %>%
    ggplot(aes(year, fertility)) +
    geom_line()
```

Code: Multiple time series

```
# line plot fertility time series for two countries- only one line (inc
countries <- c("South Korea", "Germany")
gapminder %>% filter(country %in% countries) %>%
    ggplot(aes(year, fertility)) +
    geom_line()

# line plot fertility time series for two countries - one line per coun
gapminder %>% filter(country %in% countries) %>%
    ggplot(aes(year, fertility, group = country)) +
    geom_line()

# fertility time series for two countries - lines colored by country
gapminder %>% filter(country %in% countries) %>%
    ggplot(aes(year, fertility, col = country)) +
    geom_line()
```

Code: Adding text labels to a plot

```
# life expectancy time series - lines colored by country and labeled, n
labels <- data.frame(country = countries, x = c(1975, 1965), y = c(60,
gapminder %>% filter(country %in% countries) %>%
    ggplot(aes(year, life_expectancy, col = country)) +
    geom_line() +
    geom_text(data = labels, aes(x, y, label = country), size = 5) +
    theme(legend.position = "none")
```

Learn About Verified Certificates

Stratify and Boxplot

Stratify and Boxplot

Start of transcript. Skip to the end.

0:00 / 0:00 1.0x

RAFAEL IRIZARRY: The histogram showed us that the income distribution

values show a dichotomy.

However, the histogram does not show us if the two groups of countries

are west versus the developing world.

To see distributions by geographical region.

Video

Download video file

Transcripts

Download SubRip (.srt) file

Download Text (.txt) file

Textbook link

This video corresponds to the textbook section on comparing multiple distributions with boxplots. Note that many boxplots from the video are instead dot plots in the textbook and that a different boxplot is constructed in the textbook. Also read that section to see an example of grouping factors with the `case_when` function.

Key points

- Make boxplots stratified by a categorical variable using the `geom_boxplot()` geometry.
- Rotate axis labels by changing the theme through `element_text()`. You can change the angle and justification of the text labels.
- Consider ordering your factors by a meaningful value with the `reorder()` function, which changes the order of factor levels based on a related numeric vector. This is a way to ease comparisons.
- Show the data by adding data points to the boxplot with a `geom_point()` layer. This adds information beyond the five-number summary to your plot, but too many data points it can obfuscate your message.

Code: Boxplot of GDP by region

```
# add dollars per day variable
gapminder <- gapminder %>%
    mutate(dollars_per_day = gdp/population/365)

# number of regions
length(levels(gapminder$region))

# boxplot of GDP by region in 1970
past_year <- 1970
p <- gapminder %>%
    filter(year == past_year & !is.na(gdp)) %>%
    ggplot(aes(region, dollars_per_day))
p + geom_boxplot()

# rotate names on x-axis
p + geom_boxplot() +
    theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

Code: The reorder function

```
# by default, factor order is alphabetical
fac <- factor(c("Asia", "Asia", "West", "West", "West"))
levels(fac)

# reorder factor by the category means
value <- c(10, 11, 12, 6, 4)
fac <- reorder(fac, value, FUN = mean)
levels(fac)
```

Code: Enhanced boxplot ordered by median income, scaled, and showing data

```
# reorder by median income and color by continent
p <- gapminder %>%
    filter(year == past_year & !is.na(gdp)) %>%
    mutate(region = reorder(region, dollars_per_day, FUN = median)) %>%
    ggplot(aes(region, dollars_per_day, fill = continent)) +    # color
    geom_boxplot() +
    theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
    xlab("")
p

# log2 scale y-axis
p + scale_y_continuous(trans = "log2")

# add data points
p + scale_y_continuous(trans = "log2") + geom_point(show.legend = FALSE
```

Learn About Verified Certificates

Stratify and Boxplot

Stratify and Boxplot

Start of transcript. Skip to the end.

0:00 / 0:00 1.0x

RAFAEL IRIZARRY: The histogram showed us that the income distribution

values show a dichotomy.

However, the histogram does not show us if the two groups of countries

are west versus the developing world

Video

Download video file

Transcripts

Download SubRip (.srt) file

Download Text (.txt) file

Textbook link

This video corresponds to the textbook section on comparing multiple distributions with boxplots. Note that many boxplots from the video are instead dot plots in the textbook and that a different boxplot is constructed in the textbook. Also read that section to see an example of grouping factors with the `case_when` function.

Key points

- Make boxplots stratified by a categorical variable using the `geom_boxplot()` geometry.
- Rotate axis labels by changing the theme through `element_text()`. You can change the angle and justification of the text labels.
- Consider ordering your factors by a meaningful value with the `reorder()` function, which changes the order of factor levels based on a related numeric vector. This is a way to ease comparisons.
- Show the data by adding data points to the boxplot with a `geom_point()` layer. This adds information beyond the five-number summary to your plot, but too many data points it can obfuscate your message.

Code: Boxplot of GDP by region

```
# add dollars per day variable
gapminder <- gapminder %>%
    mutate(dollars_per_day = gdp/population/365)

# number of regions
length(levels(gapminder$region))

# boxplot of GDP by region in 1970
past_year <- 1970
p <- gapminder %>%
    filter(year == past_year & !is.na(gdp)) %>%
    ggplot(aes(region, dollars_per_day))
p + geom_boxplot()

# rotate names on x-axis
p + geom_boxplot() +
    theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

Code: The reorder function

```
# by default, factor order is alphabetical
fac <- factor(c("Asia", "Asia", "West", "West", "West"))
levels(fac)

# reorder factor by the category means
value <- c(10, 11, 12, 6, 4)
fac <- reorder(fac, value, FUN = mean)
levels(fac)
```

Code: Enhanced boxplot ordered by median income, scaled, and showing data

```
# reorder by median income and color by continent
p <- gapminder %>%
    filter(year == past_year & !is.na(gdp)) %>%
    mutate(region = reorder(region, dollars_per_day, FUN = median)) %>%
    ggplot(aes(region, dollars_per_day, fill = continent)) +    # color
    geom_boxplot() +
    theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
    xlab("")
p

# log2 scale y-axis
p + scale_y_continuous(trans = "log2")

# add data points
p + scale_y_continuous(trans = "log2") + geom_point(show.legend = FALSE
```

Learn About Verified Certificates

Comparing Distributions

Comparing Distributions

Start of transcript. Skip to the end.

0:00 / 0:00 1.0x

The exploratory data analysis we have conducted

has revealed two characteristics about average income distributions in 1970.

Using a histogram, we found a bimodal distribution

with the most relating to poor and rich countries

Video

Download video file

Transcripts

Download SubRip (.srt) file

Download Text (.txt) file

Textbook link

This video corresponds to the textbook section on 1970 versus 2010 income distributions. Note that the boxplots are slightly different: the group variable in those plots was defined in section 10.7.1.

Key points

- Use `intersect()` to find the overlap between two vectors.
- To make boxplots where grouped variables are adjacaent, color the boxplot by a factor instead of faceting by that factor. This is a way to ease comparisons.
- The data suggest that the income gap between rich and poor countries has narrowed, not expanded.

Code: Histogram of income in West versus developing world, 1970 and 2010

```
# add dollars per day variable and define past year
gapminder <- gapminder %>%
    mutate(dollars_per_day = gdp/population/365)
past_year <- 1970

# define Western countries
west <- c("Western Europe", "Northern Europe", "Southern Europe", "Nort

# facet by West vs devloping
gapminder %>%
    filter(year == past_year & !is.na(gdp)) %>%
    mutate(group = ifelse(region %in% west, "West", "Developing")) %>%
    ggplot(aes(dollars_per_day)) +
    geom_histogram(binwidth = 1, color = "black") +
    scale_x_continuous(trans = "log2") +
    facet_grid(. ~ group)

# facet by West/developing and year
present_year <- 2010
gapminder %>%
    filter(year %in% c(past_year, present_year) & !is.na(gdp)) %>%
    mutate(group = ifelse(region %in% west, "West", "Developing")) %>%
    ggplot(aes(dollars_per_day)) +
    geom_histogram(binwidth = 1, color = "black") +
    scale_x_continuous(trans = "log2") +
    facet_grid(year ~ group)
```

Code: Income distribution of West versus developing world, only countries with data

```
# define countries that have data available in both years
country_list_1 <- gapminder %>%
    filter(year == past_year & !is.na(dollars_per_day)) %>% .$country
    country_list_2 <- gapminder %>%
    filter(year == present_year & !is.na(dollars_per_day)) %>% .$countr
    country_list <- intersect(country_list_1, country_list_2)

# make histogram including only countries with data available in both y
gapminder %>%
    filter(year %in% c(past_year, present_year) & country %in% country_
    mutate(group = ifelse(region %in% west, "West", "Developing")) %>%
    ggplot(aes(dollars_per_day)) +
    geom_histogram(binwidth = 1, color = "black") +
    scale_x_continuous(trans = "log2") +
    facet_grid(year ~ group)
```

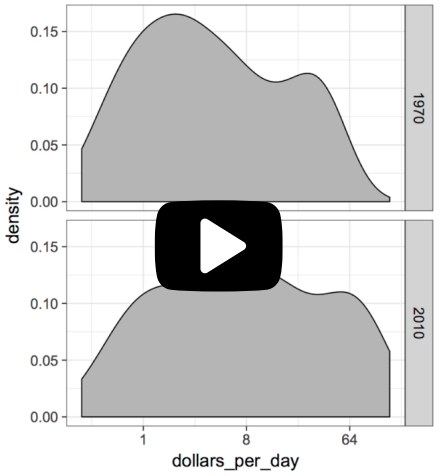
Code: Boxplots of income in West versus developing world, 1970 and 2010

```
p <- gapminder %>%
    filter(year %in% c(past_year, present_year) & country %in% country_
    mutate(region = reorder(region, dollars_per_day, FUN = median)) %>%
    ggplot() +
    theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
    xlab("") + scale_y_continuous(trans = "log2")

 p + geom_boxplot(aes(region, dollars_per_day, fill = continent)) +
     facet_grid(year ~ .)

 # arrange matching boxplots next to each other, colored by year
 p + geom_boxplot(aes(region, dollars_per_day, fill = factor(year)))
```

Learn About Verified Certificates

Density Plots

Density Plots



0:00 / 0:00 1.0x

Start of transcript. Skip to the end.

RAFAEL IRIZARRY: We have used data exploration

to discover that the income gap between rich and poor countries

has closed considerably during the last forty years.

We use a series of histograms and box plots to see this.

Video

Download video file

Transcripts

Download SubRip (.srt) file

Download Text (.txt) file

Textbook link

This video corresponds to the following sections:

- The end of the textbook section on 1970 versus 2010 income distributions
- Textbook section on accessing computed variables
- Textbook section on weighted densities

Key points

- Change the y-axis of density plots to variable counts using `..count..` as the y argument.
- The `case_when()` function defines a factor whose levels are defined by a variety of logical operations to group data.
- Plot stacked density plots using `position="stack"`.
- Define a weight aesthetic mapping to change the relative weights of density plots - for example, this allows weighting of plots by population rather than number of countries.

Code: Faceted smooth density plots

```
# see the code below the previous video for variable definitions

# smooth density plots - area under each curve adds to 1
gapminder %>%
    filter(year == past_year & country %in% country_list) %>%
    mutate(group = ifelse(region %in% west, "West", "Developing")) %>%
    summarize(n = n()) %>% knitr::kable()

# smooth density plots - variable counts on y-axis
p <- gapminder %>%
    filter(year == past_year & country %in% country_list) %>%
    mutate(group = ifelse(region %in% west, "West", "Developing")) %>%
    ggplot(aes(dollars_per_day, y = ..count.., fill = group)) +
    scale_x_continuous(trans = "log2")
p + geom_density(alpha = 0.2, bw = 0.75) + facet_grid(year ~ .)
```

Code: Add new region groups with case_when

```
# add group as a factor, grouping regions
gapminder <- gapminder %>%
    mutate(group = case_when(
            .$region %in% west ~ "West",
            .$region %in% c("Eastern Asia", "South-Eastern Asia") ~ "Ea
            .$region %in% c("Caribbean", "Central America", "South Amer
            .$continent == "Africa" & .$region != "Northern Africa" ~ "
            TRUE ~ "Others"))

# reorder factor levels
gapminder <- gapminder %>%
    mutate(group = factor(group, levels = c("Others", "Latin America",
```

Code: Stacked density plot

```
# note you must redefine p with the new gapminder object first
p <- gapminder %>%
  filter(year %in% c(past_year, present_year) & country %in% country_li
    ggplot(aes(dollars_per_day, fill = group)) +
    scale_x_continuous(trans = "log2")

# stacked density plot
p + geom_density(alpha = 0.2, bw = 0.75, position = "stack") +
    facet_grid(year ~ .)
```

Code: Weighted stacked density plot

```
# weighted stacked density plot
gapminder %>%
    filter(year %in% c(past_year, present_year) & country %in% country_
    group_by(year) %>%
    mutate(weight = population/sum(population*2)) %>%
    ungroup() %>%
    ggplot(aes(dollars_per_day, fill = group, weight = weight)) +
    scale_x_continuous(trans = "log2") +
    geom_density(alpha = 0.2, bw = 0.75, position = "stack") + facet_gr
```

Learn About Verified Certificates

Ecological Fallacy

Ecological Fallacy

Start of transcript. Skip to the end.

0:00 / 0:00 1.0x

RAFAEL IRIZARRY: Throughout this section,

we have been comparing regions of the world.

We have seen that on average some regions do better

than others in health outcomes and economic outcomes.

Video

Download video file

Transcripts

Download SubRip (.srt) file

Download Text (.txt) file

Textbook link

This video corresponds to the textbook section on the ecological fallacy.

Key points

- The `breaks` argument allows us to set the location of the axis labels and tick marks.
- The *logistic* or *logit transformation* is defined as $f(p) = \log\frac{p}{1-p}$, or the log of odds. This scale is useful for highlighting differences near 0 or near 1 and converts fold changes into constant increases.
- The *ecological fallacy* is assuming that conclusions made from the average of a group apply to all members of that group.

Code

```
# define gapminder
library(tidyverse)
library(dslabs)
data(gapminder)

# add additional cases
gapminder <- gapminder %>%
    mutate(group = case_when(
        .$region %in% west ~ "The West",
        .$region %in% "Northern Africa" ~ "Northern Africa",
        .$region %in% c("Eastern Asia", "South-Eastern Asia") ~ "East A
        .$region == "Southern Asia" ~ "Southern Asia",
        .$region %in% c("Central America", "South America", "Caribbean"
        .$continent == "Africa" & .$region != "Northern Africa" ~ "Sub-
        .$region %in% c("Melanesia", "Micronesia", "Polynesia") ~ "Paci

# define a data frame with group average income and average infant surv
surv_income <- gapminder %>%
    filter(year %in% present_year & !is.na(gdp) & !is.na(infant_mortali
    group_by(group) %>%
    summarize(income = sum(gdp)/sum(population)/365,
                         infant_survival_rate = 1 - sum(infant_mortality
surv_income %>% arrange(income)

# plot infant survival versus income, with transformed axes
surv_income %>% ggplot(aes(income, infant_survival_rate, label = group,
    scale_x_continuous(trans = "log2", limit = c(0.25, 150)) +
    scale_y_continuous(trans = "logit", limit = c(0.875, .9981),
                                        breaks = c(.85, .90, .95, .99, .
    geom_label(size = 3, show.legend = FALSE)
```

Learn About Verified Certificates