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