Package ‘lvplot’

August 29, 2016

Version 0.2.0
Title Letter Value 'Boxplots'
Description Implements the letter value 'boxplot' which extends the standard
'boxplot' to deal with both larger and smaller number of data points
by dynamically selecting the appropriate number of letter values to display.
License GPL (>= 2)
Depends R (>= 2.14)
Imports ggplot2 (>= 2.0.0), grid, RColorBrewer
Suggests depth
RoxygenNote 5.0.1
LazyData true
LazyDataCompression xz
NeedsCompilation no
Author Hadley Wickham [aut, cre],
Heike Hofmann [aut]
Maintainer Hadley Wickham <hadley@rstudio.com>
Repository CRAN
Date/Publication 2016-05-01 18:56:01

R topics documented:
census ................................................. 2
determineDepth ...................................... 2
geom_Lv ................................................ 3
LVboxplot ............................................ 5
lvtable ............................................... 7
ontime .............................................. 7

Index 9
census  County demographics based on 1980 US Census

Description
County level statistics based on the 1980 US Census.

Usage
census

Format
A data frame with 10 variables
- county  County name
- FIPS  FIPS county code
- Latitude,Longitude  Geographic location of county centers
- JanTmp,JulTmp  (normalized) Temperatures in January & July
- JanSun,JulSun  (normalized) Sunshine measurement in January & July
- Elevtn  Elevation above sea level
- totalpop  Population

determineDepth  Determine depth of letter values needed for n observations.

Description
Determine depth of letter values needed for n observations.

Usage
determineDepth(n, k = NULL, alpha = NULL, perc = NULL)

Arguments
- n  number of observation to be shown in the LV boxplot
- k  number of letter value statistics used
- alpha  if supplied, depth k is calculated such that (1-alpha)100 intervals of an LV statistic do not extend into neighboring LV statistics.
- perc  if supplied, depth k is adjusted such that perc percent outliers are shown

Details
Supply one of k, alpha or perc.
Side-by-side LV boxplots with ggplot2.

Description
An extension of standard boxplots which draws k letter statistics. Conventional boxplots (Tukey 1977) are useful displays for conveying rough information about the central 50% of the data and the extent of the data. For moderate-sized data sets (n < 1000), detailed estimates of tail behavior beyond the quartiles may not be trustworthy, so the information provided by boxplots is appropriately somewhat vague beyond the quartiles, and the expected number of “outliers” and “far-out” values for a Gaussian sample of size n is often less than 10 (Hoaglin, Iglewicz, and Tukey 1986). Large data sets (n ≈ 10,000 – 100,000) afford more precise estimates of quantiles in the tails beyond the quartiles and also can be expected to present a large number of “outliers” (about 0.4 + 0.007n). The letter-value box plot addresses both these shortcomings: it conveys more detailed information in the tails using letter values, only out to the depths where the letter values are reliable estimates of their corresponding quantiles (corresponding to tail areas of roughly 2^{-i}); “outliers” are defined as a function of the most extreme letter value shown. All aspects shown on the letter-value boxplot are actual observations, thus remaining faithful to the principles that governed Tukey's original boxplot.

Usage

```r
geom_lv(mapping = NULL, data = NULL, stat = "lv", position = "dodge",
    outlier.colour = "black", outlier.shape = 19, outlier.size = 1.5,
    outlier.stroke = 0.5, na.rm = TRUE, varwidth = FALSE,
    width.method = "linear", show.legend = NA, inherit.aes = TRUE, ...)
```

```r
GeomLv
```

```r
scale_fill_lv(...)
```

```r
stat_lv(mapping = NULL, data = NULL, geom = "lv", position = "dodge",
    na.rm = TRUE, conf = 0.95, percent = NULL, k = NULL,
    show.legend = NA, inherit.aes = TRUE, ...)
```

```r
StatLv
```

Arguments

- **mapping**: Set of aesthetic mappings created by `aes` or `aes_`. If specified and `inherit.aes = TRUE` (the default), it is combined with the default mapping at the top level of the plot. You must supply `mapping` if there is no plot mapping.

- **data**: The data to be displayed in this layer. There are three options:
  - If `NULL`, the default, the data is inherited from the plot data as specified in the call to `ggplot`.
  - A `data.frame`, or other object, will override the plot data. All objects will be fortified to produce a data frame. See `fortify` for which variables will be created.
A function will be called with a single argument, the plot data. The return value must be a data.frame, and will be used as the layer data.

position
Position adjustment, either as a string, or the result of a call to a position adjustment function.

outlier.colour
Override aesthetics used for the outliers. Defaults come from geom_point().

outlier.shape
Override aesthetics used for the outliers. Defaults come from geom_point().

outlier.size
Override aesthetics used for the outliers. Defaults come from geom_point().

outlier.stroke
Override aesthetics used for the outliers. Defaults come from geom_point().

na.rm
If FALSE (the default), removes missing values with a warning. If TRUE silently removes missing values.

varwidth
if FALSE (default) draw boxes that are the same size for each group. If TRUE, boxes are drawn with widths proportional to the square-roots of the number of observations in the groups (possibly weighted, using the weight aesthetic).

width.method
character, one of 'linear' (default), 'area', or 'height'. This parameter determines whether the width of the box for letter value LV(i) should be proportional to i (linear), proportional to $2^{-i}$ (height), or whether the area of the box should be proportional to $2^{-i}$ (area).

show.legend
logical. Should this layer be included in the legends? NA, the default, includes if any aesthetics are mapped. FALSE never includes, and TRUE always includes.

inherit.aes
If FALSE, overrides the default aesthetics, rather than combining with them. This is most useful for helper functions that define both data and aesthetics and shouldn’t inherit behaviour from the default plot specification, e.g. borders.

... other arguments passed on to layer. These are often aesthetics, used to set an aesthetic to a fixed value, like color = "red" or size = 3. They may also be parameters to the paired geom/stat.

gem, stat
Use to override the default connection between geom_lv and stat_lv.

conf
confidence level

percent
numeric value: percent of data in outliers

k
number of letter values shown

Format
An object of class GeomLv (inherits from Geom, ggproto) of length 6.

Computed-reported variables

k Number of Letter Values used for the display
LV Name of the Letter Value
width width of the interquartile box

References
LVboxplot

See Also

`stat_quantile` to view quantiles conditioned on a continuous variable.

Examples

```r
library(ggplot2)
p <- ggplot(mpg, aes(class, hwy))
p + geom_lv(aes(fill = ..LV..)) + scale_fill_brewer()
p + geom_lv() + geom_jitter(width = 0.2)
p + geom_lv(alpha = 1, aes(fill = ..LV..)) + scale_fill_lv()

# Outliers
p + geom_lv(varwidth = TRUE, aes(fill = ..LV..)) + scale_fill_lv()
p + geom_lv(fill = "grey80", colour = "black")
p + geom_lv(outlier.colour = "red", outlier.shape = 1)

# Plots are automatically dodged when any aesthetic is a factor
p + geom_lv(aes(fill = drv))

# varwidth adjusts the width of the boxes according to the number of observations
p <- ggplot(ontime, aes(uniqueCarrier, TaxiIn + TaxiOut)) +
  geom_lv(aes(fill = ..LV..), varwidth=TRUE) +
  scale_fill_lv() +
  scale_y_sqrt() +
  theme_bw()

ontime$DayOfWeek <- as.POSIXlt(ontime$FlightDate)$wday
p <- ggplot(ontime, aes(factor(DayOfWeek), TaxiIn + TaxiOut)) +
  geom_lv(aes(fill = ..LV..)) +
  scale_fill_lv() +
  scale_y_sqrt() +
  theme_bw()
```

LVboxplot

Side-by-side LV boxplots with base graphics

Description

An extension of standard boxplots which draws k letter statistics. Conventional boxplots (Tukey 1977) are useful displays for conveying rough information about the central 50% of the data and the extent of the data.

Usage

```r
LVboxplot(x, ...)
```

## S3 method for class 'formula'
```r
LVboxplot(formula, alpha = 0.95, k = NULL, perc = NULL,
  horizontal = TRUE, xlab = NULL, ylab = NULL, col = "grey30",
```
bg = "grey90", width = 0.9, width.method = "linear",
median.col = "grey10", ...)"

## S3 method for class 'numeric'
LVboxplot(x, alpha = 0.95, k = NULL, perc = NULL,
   horizontal = TRUE, xlab = NULL, ylab = NULL, col = "grey90",
   bg = "grey90", width = 0.9, width.method = "linear",
   median.col = "grey10", ...)"

**Arguments**

- **x**: numeric vector of data
- **...**: passed onto `plot`
- **formula**: a plotting formula of the form `y ~ x`, where `x` is a string or factor. The values of `y` will be split into groups according to their values on `x` and separate letter value box plots of `y` are drawn side by side in the same display.
- **alpha**: if supplied, depth `k` is calculated such that `(1-alpha)100` intervals of an LV statistic do not extend into neighboring LV statistics.
- **k**: number of letter value statistics used
- **perc**: if supplied, depth `k` is adjusted such that `perc` percent outliers are shown
- **horizontal**: display horizontally (TRUE) or vertically (FALSE)
- **xlab**: x axis label
- **ylab**: y axis label
- **col**: vector of colours to use
- **bg**: background colour
- **width**: maximum height/width of box
- **width.method**: one of 'linear', 'height' or 'area'. Methods 'height' and 'area' ensure that these dimension are proportional to the number of observations within each box.
- **median.col**: colour of the line for the median

**Details**

For moderate-sized data sets (`n < 1000`), detailed estimates of tail behavior beyond the quartiles may not be trustworthy, so the information provided by boxplots is appropriately somewhat vague beyond the quartiles, and the expected number of “outliers” and “far-out” values for a Gaussian sample of size `n` is often less than 10 (Hoaglin, Iglewicz, and Tukey 1986). Large data sets (`n ≈ 10,000 – 100,000`) afford more precise estimates of quantiles in the tails beyond the quartiles and also can be expected to present a large number of “outliers” (about `0.4 + 0.007n`).

The letter-value box plot addresses both these shortcomings: it conveys more detailed information in the tails using letter values, only out to the depths where the letter values are reliable estimates of their corresponding quantiles (corresponding to tail areas of roughly `2^−i`); “outliers” are defined as a function of the most extreme letter value shown. All aspects shown on the letter-value boxplot are actual observations, thus remaining faithful to the principles that governed Tukey’s original boxplot.
Examples

```r
n <- 10
oldpar <- par()
par(mfrow=c(4,2), mar=c(3,3,3,3))
for (i in 1:4) {
  x <- rexp(10 ^ (i + 1))
  boxplot(x, col = "grey", horizontal = TRUE)
  title(paste("Exponential, n = ", length(x)))
  LVboxplot(x, col = "grey", xlab = "")
}
par(mfrow=oldpar$mfrow, mar=oldpar$mar)

with(ontime, LVboxplot(sqrt(TaxiIn + TaxiOut) ~ UniqueCarrier, horizontal=FALSE))
```

---

**lvtbl**

*Compute table of k letter values for vector x*

**Description**

Compute table of k letter values for vector x

**Usage**

```r
lvtbl(x, k, alpha = 0.95)
```

**Arguments**

- `x`: input numeric vector
- `k`: number of letter values to compute
- `alpha`: alpha-threshold for confidence level

---

**ontime**

*Ontime Flight Data*

**Description**

Data set detailing on-time performance of national US flights in January 2015. This data is a subset of the data provided by the US Department of Transportation. The full data as well as archived or more recent data is available for download from [http://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236&DB_Short_Name=On-Time](http://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236&DB_Short_Name=On-Time).

**Usage**

```r
ontime
```
Format

A data frame consisting of the variables

- **FlightDate** a date variable of the day of the flight
- **UniqueCarrier** factor variable of the carrier (using the two letter abbreviation)
- **FlightNum** numeric variable of the flight number
- **CRSDepTime** scheduled departure time in hhmm format
- **DepTime** actual departure time in hhmm format
- **CRSArrTime** scheduled arrival time in hhmm format
- **ArrTime** actual arrival time in hhmm format
- **TaxiOut** numeric variable of the taxi out time in minutes
- **TaxiIn** numeric variable of the taxi in time in minutes
- **ArrDelay** Arrival delay, in Minutes
- **DepDelay** Departure delay, in Minutes
- **CarrierDelay** Carrier Delay, in Minutes
- **WeatherDelay** Weather Delay, in Minutes
- **NASDelay** National Air System Delay, in Minutes
- **SecurityDelay** Security Delay, in Minutes
- **LateAircraftDelay** Late Aircraft Delay, in Minutes

References

http://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236&DB_Short_Name=On-Time

Examples

```r
library(ggplot2)
ggplot(ontime, aes(UniqueCarrier, TaxiIn + TaxiOut)) +
  geom_lw(aes(fill = ..LV..)) +
  scale_fill_lw() +
  scale_y_sqrt() +
  theme_bw()
```
Index

*Topic datasets
  - census, 2
  - geom_lv, 3
  - ontime, 7

aes, 3
aes_, 3
borders, 4
census, 2
determineDepth, 2
fortify, 3
geom_lv, 3
GeomLv(geom_lv), 3
ggplot, 3
layer, 4
LVboxplot, 5
lvtable, 7
ontime, 7
plot, 6
scale_fill_lv(geom_lv), 3
stat_lv(geom_lv), 3
stat_quantile, 5
StatLv(geom_lv), 3