Continuous Variables, pt. 2

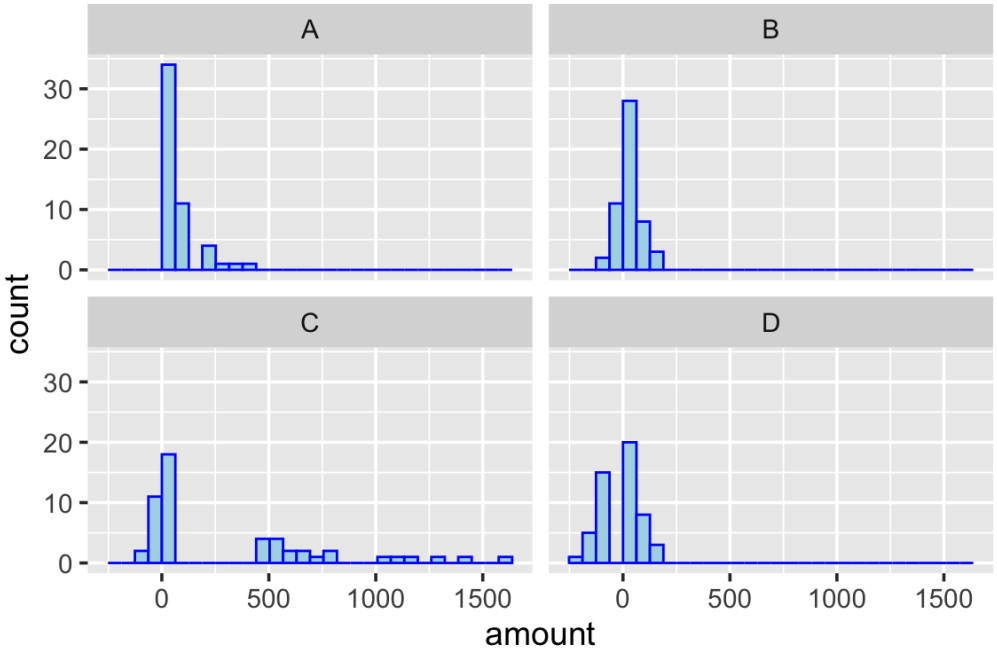
Weekly Savings

A	B	C	D
0.909	-95.249	-95.249	-195.2
1.496	-75.119	-75.119	-175.1
2.024	-61.774	-61.774	-161.8
2.405	-46.180	-46.180	-146.2
3.272	-39.818	-39.818	-139.8
4.769	-37.617	-37.617	-137.6
5.578	-22.619	-22.619	-122.6
6.647	-16.224	-16.224	-116.2
7.927	-6.186	-6.186	-106.2
10.859	-4.291	-4.291	-104.3
12.035	-3.251	-3.251	-103.3
13.924	-2.036	-2.036	-102.0
14.074	-1.523	-1.523	-101.5
14.227	0.577	0.577	-99.4
14.580	8.376	8.376	-91.6
16.234	10.078	10.078	-89.9
18.852	13.598	13.598	-86.4
19.980	16.915	16.915	-83.1

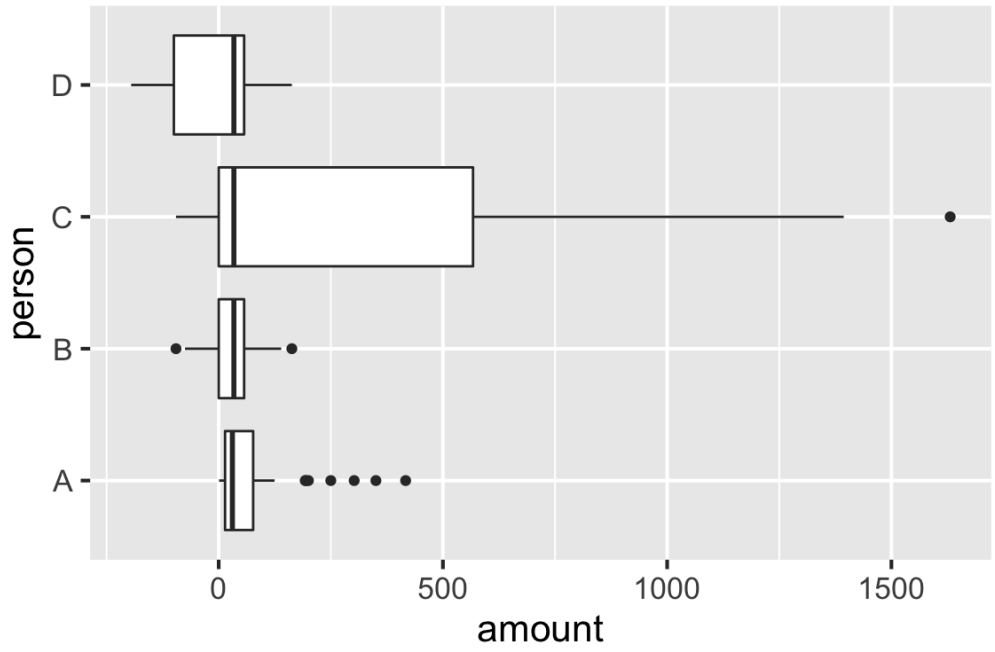
A	B	C	D
24.440	17.470	17.470	-82.5
25.109	18.648	18.648	-81.4
25.676	20.098	20.098	-79.9
25.867	24.333	24.333	24.3
26.000	28.198	28.198	28.2
28.535	31.104	31.104	31.1
29.543	31.805	31.805	31.8
30.478	32.744	32.744	32.7
30.648	35.035	35.035	35.0
39.095	37.773	37.773	37.8
40.210	40.510	40.510	40.5
47.266	40.707	40.707	40.7
51.398	41.001	41.001	41.0
52.306	45.793	457.933	45.8
57.083	48.475	484.753	48.5
58.269	49.300	493.004	49.3
65.167	49.784	497.842	49.8
65.548	52.184	521.844	52.2
73.493	52.619	526.191	52.6
73.726	54.153	541.527	54.2

A	B	C	D
74.934	55.683	556.831	55.7
82.537	59.798	597.985	59.8
85.918	62.602	626.023	62.6
92.275	65.141	651.414	65.1
95.689	65.371	653.706	65.4
104.578	70.356	703.560	70.4
124.599	76.699	766.989	76.7
192.962	78.215	782.148	78.2
194.340	103.500	1034.998	103.5
199.995	109.222	1092.217	109.2
249.964	119.499	1194.992	119.5
302.121	128.147	1281.472	128.1
350.536	139.366	1393.657	139.4
416.852	163.109	1631.089	163.1

Histograms

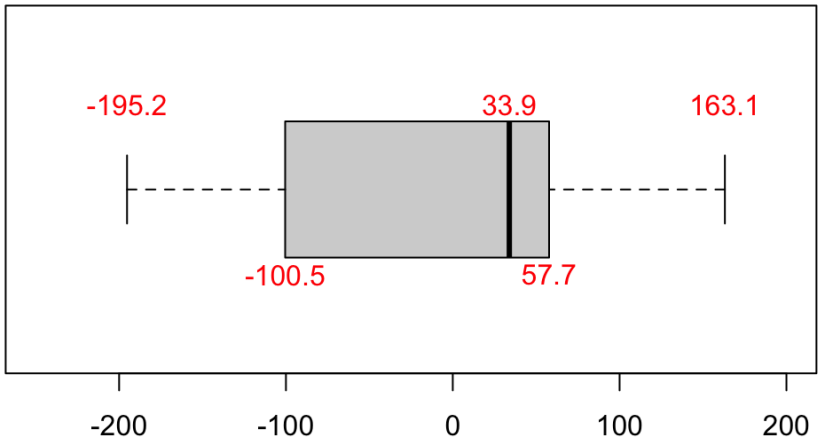


Boxplots - difficult to show bimodality



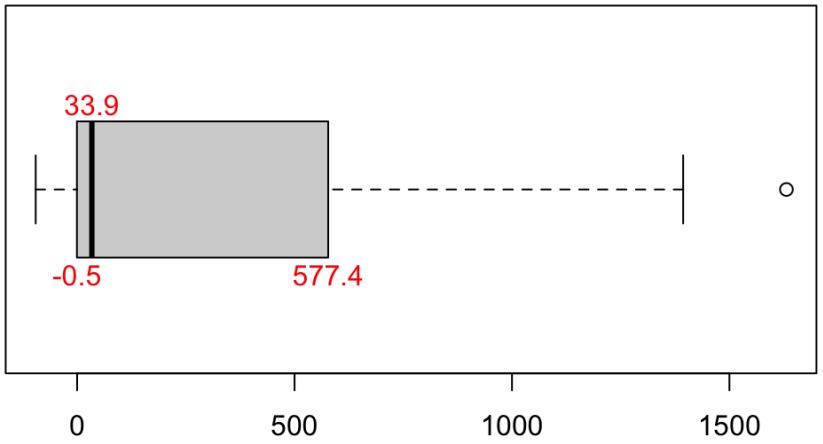
Boxplot (Person “D”)

```
##         min lower-hinge      median upper-hinge         max
##      -195.2      -100.5        33.9        57.7       163.1
```



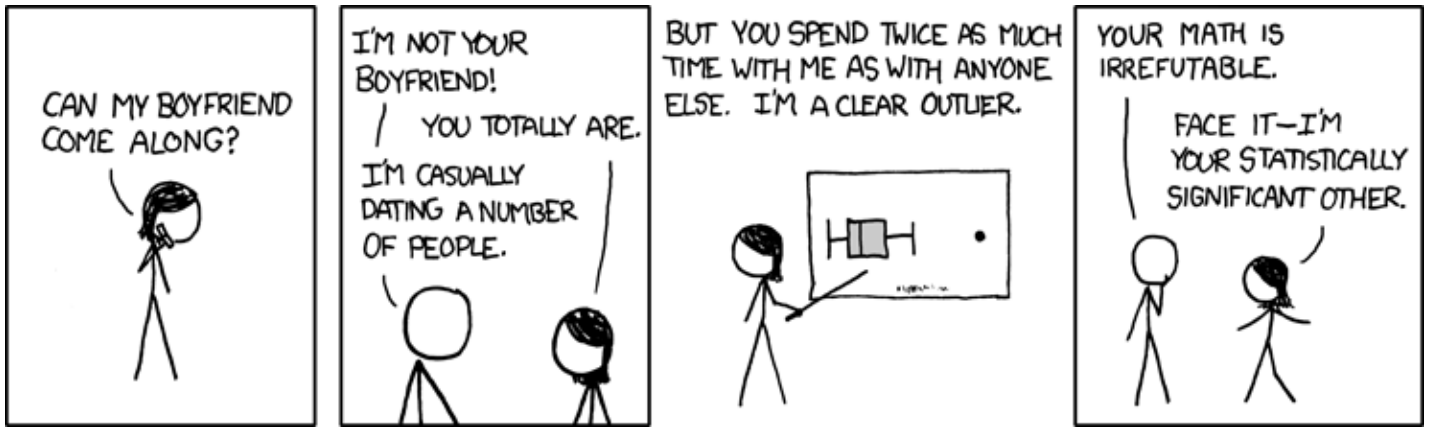
Boxplot with outliers (Person “C”)

```
##         min lower-hinge      median upper-hinge         max
##     -95.249      -0.473      33.889     577.408    1631.089
```



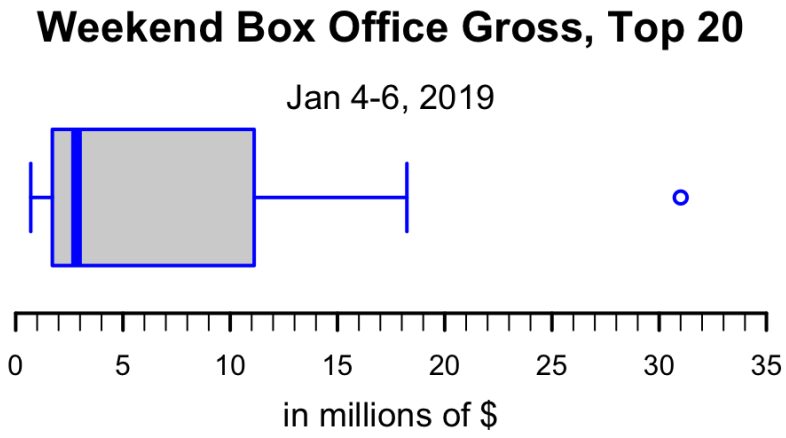
What does it take to be an outlier?

What does it take to be an outlier?



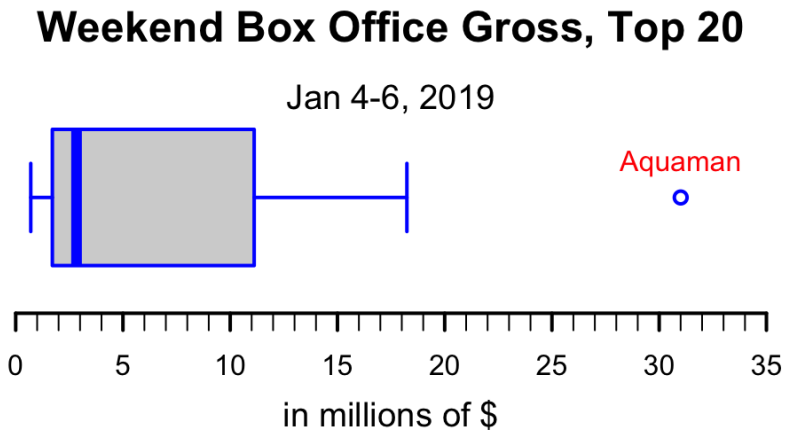
https://www.explainxkcd.com/wiki/index.php/539:_Boyfriend

What does it take to be an outlier?

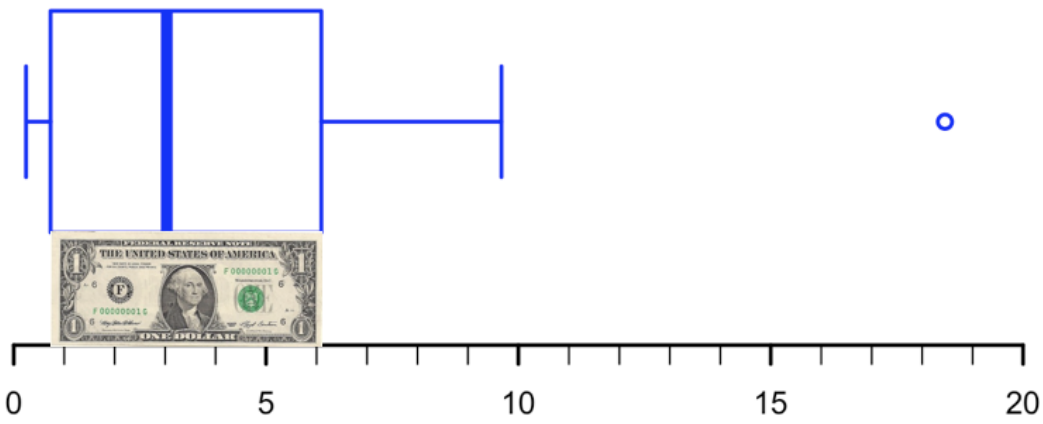


Source: http://www.boxofficemojo.com/weekend/chart/

What does it take to be an outlier?



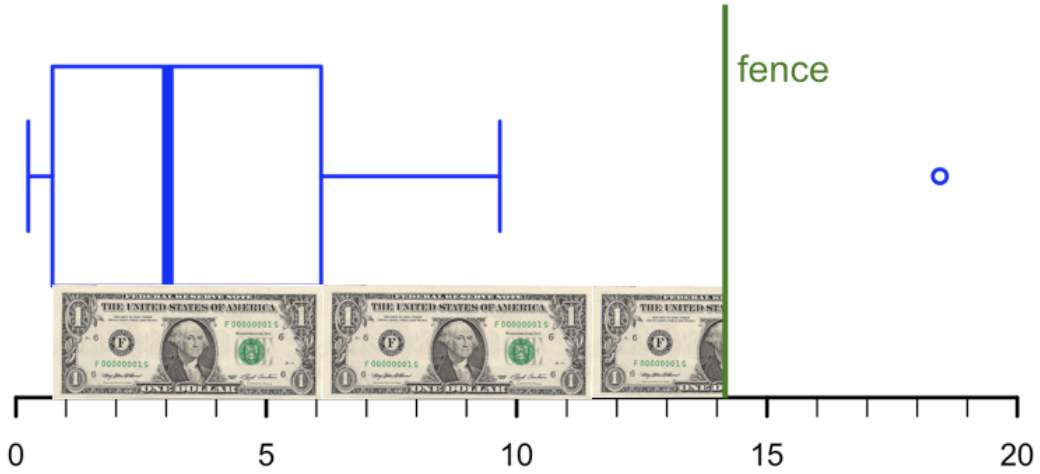
What does it take to be an outlier?



Q3 - Q4

“H-spread” or fourth spread (upper hinge - lower hinge)

What does it take to be an outlier?

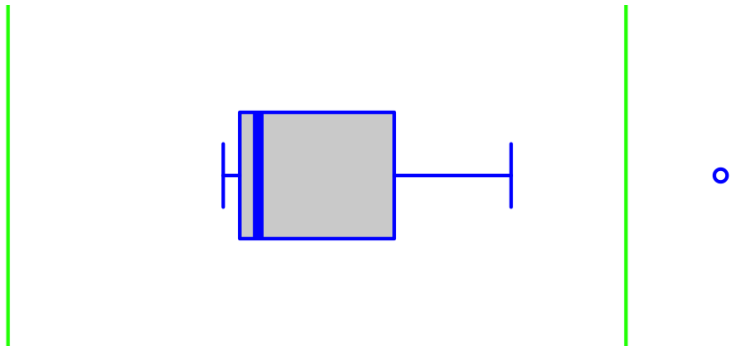


fences:

1.5 x hinge or fourth spread above upper-hinge

1.5 x hinge or fourth spread below lower-hinge

Fences

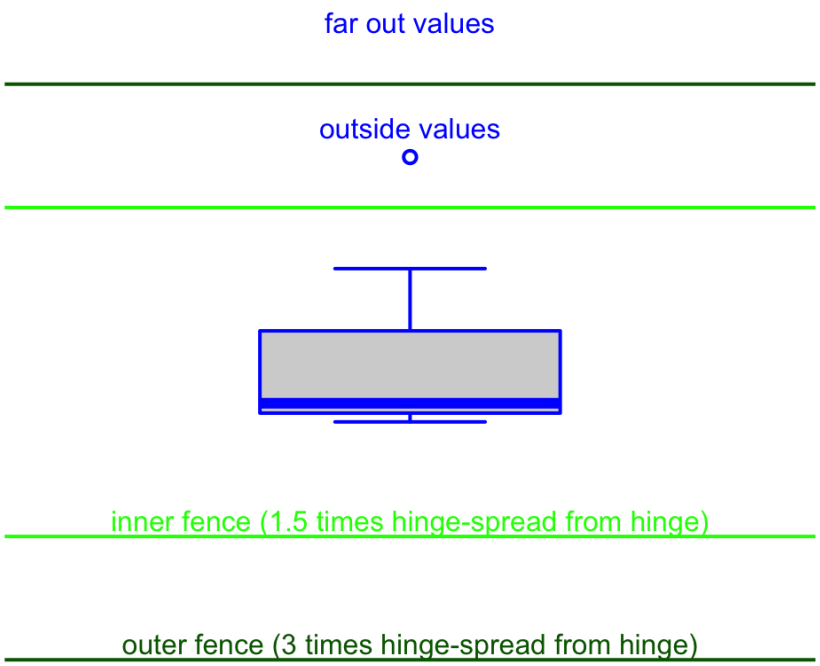


fences:

1.5 x hinge or fourth spread above upper-hinge

1.5 x hinge or fourth spread below lower-hinge

Tukey’s original boxplot



Quartiles

```
##  [1]  0.703  0.923  1.005  1.168  1.609  1.808  1.843  1.903  2.147  2.368
## [11]  3.303  4.674  4.755  5.735  9.110 13.127 13.203 15.861 18.238 31.003
```

```
##         min lower-hinge      median upper-hinge         max
##       0.703       1.709       2.835      11.118      31.003
```

```
##     0%    25%    50%    75%   100%
##  0.703  1.758  2.835 10.114 31.003
```

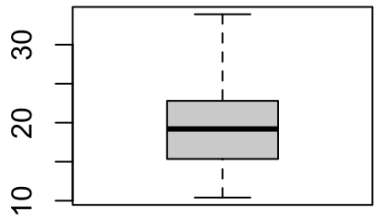
See: ?quantile for different methods

Sometimes boxplots are drawn using the IQR (interquartile range) instead of hinge spread

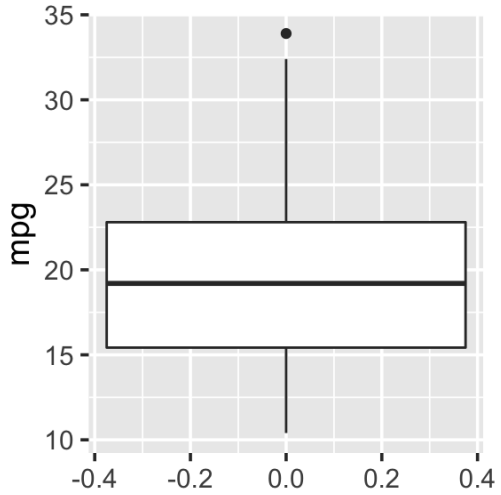
lower-hinge is either the lowest number on the dataset or the average?

base R vs. ggplot2

```
boxplot(mtcars$mpg)
```




```
ggplot(mtcars, aes(y = mpg)) +
    geom_boxplot() +
    theme_grey(14)
```



"take these outliers with a grain of salt'

boxplot stats

```
# base R
boxplot.stats(mtcars$mpg)
```

```
## $stats
## [1] 10.4 15.3 19.2 22.8 33.9
##
## $n
## [1] 32
##
## $conf
## [1] 17.1 21.3
##
## $out
## numeric(0)
```

```
# ggplot2
g <- ggplot(mtcars, aes(y = mpg)) +
    geom_boxplot()
ggplot_build(g)$data[[1]][,1:6]
```

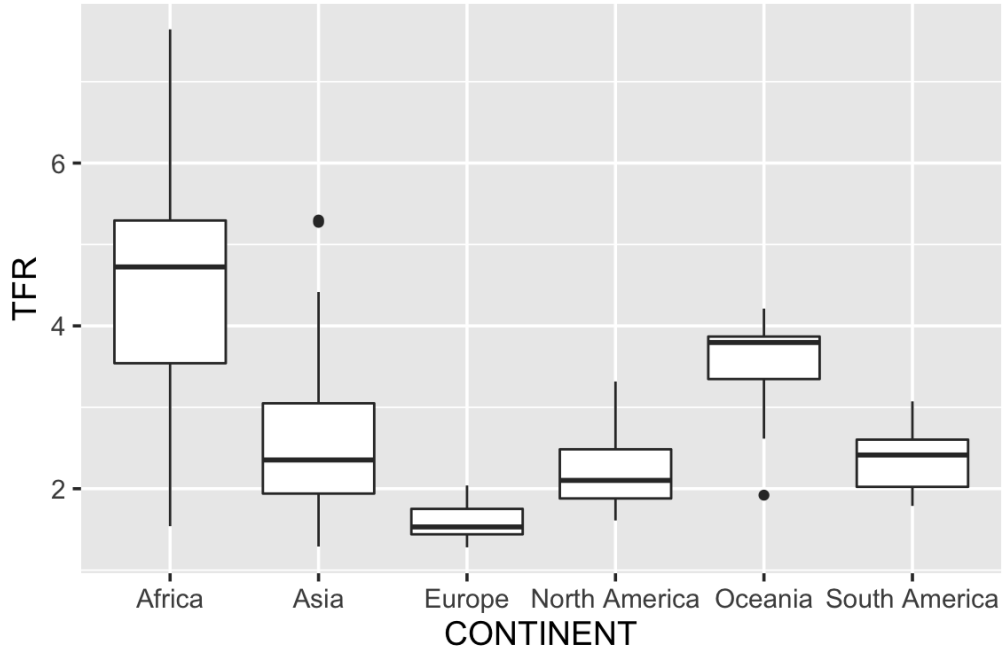
ymin	lower	middle	upper	ymax	outliers
10.4	15.4	19.2	22.8	32.4	33.9

= five number summary

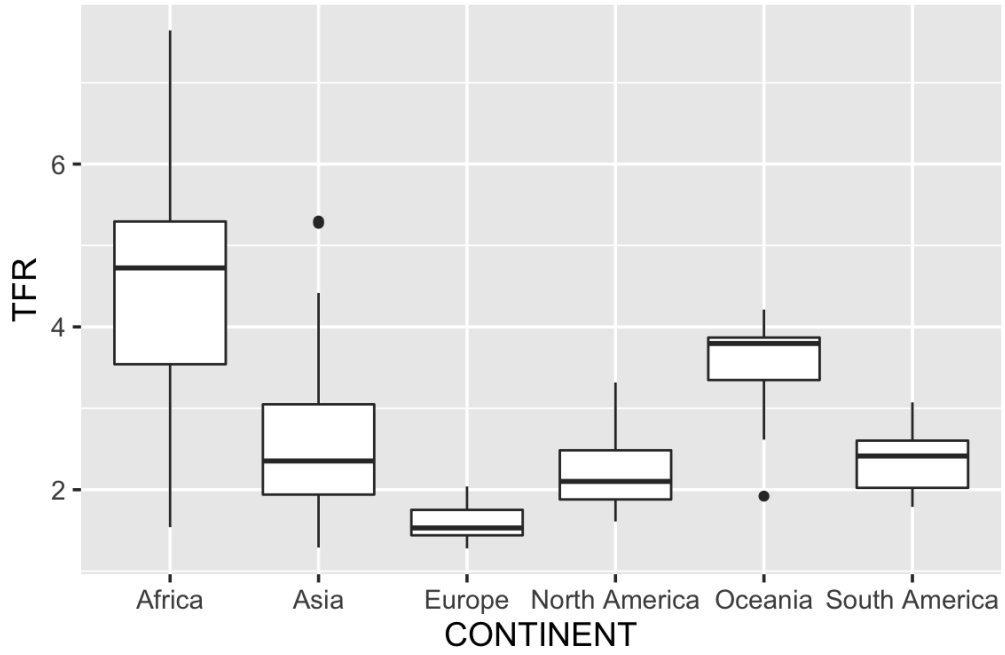
```
quantile(mtcars$mpg)
```

```
##   0%  25%  50%  75% 100%
## 10.4 15.4 19.2 22.8 33.9
```

Multiple boxplots

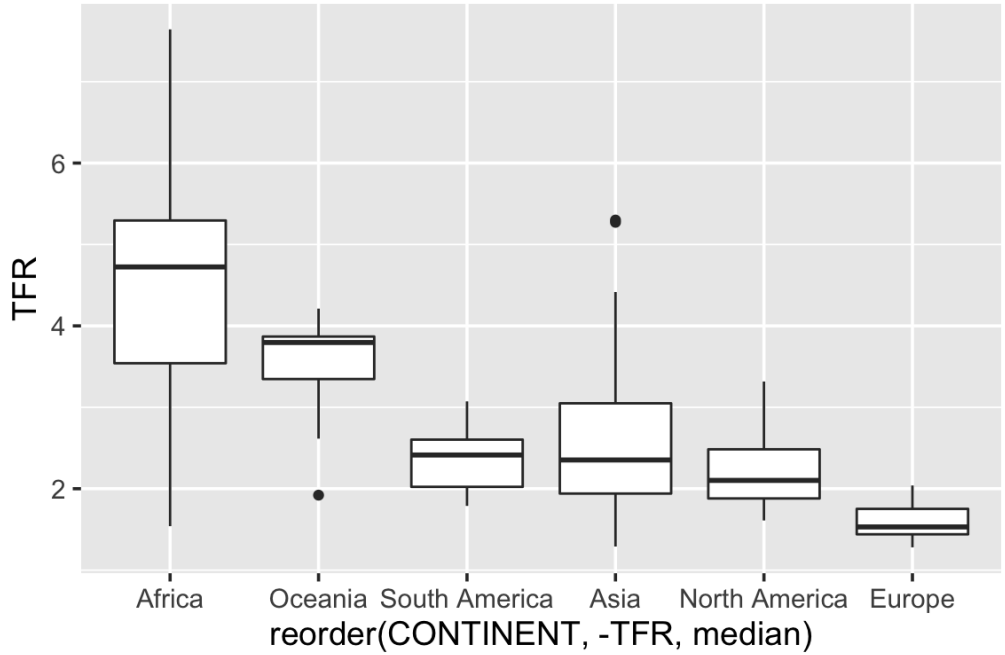


Multiple boxplots



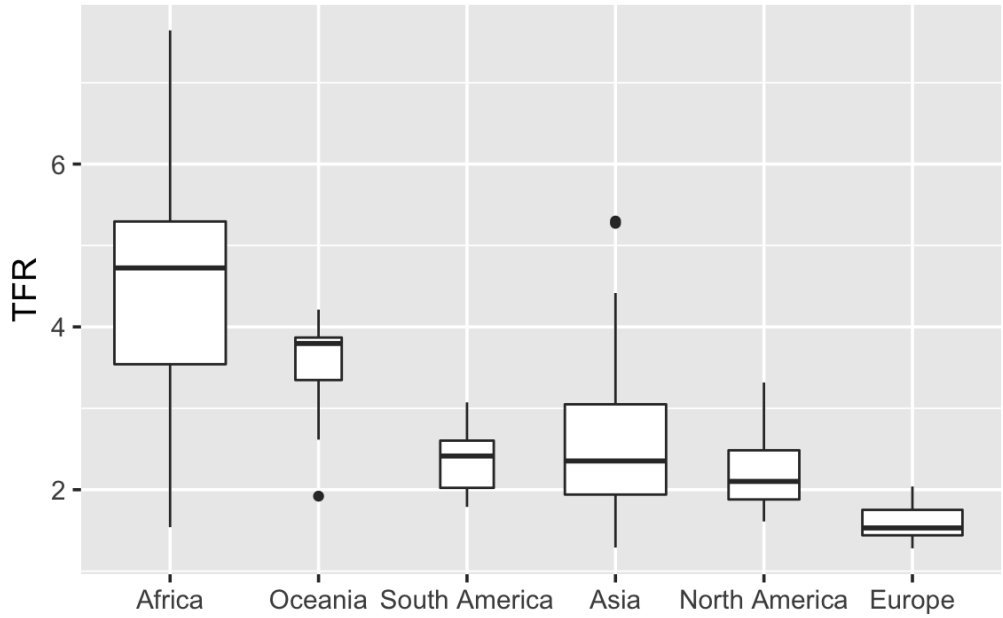
COUNTRY	CONTINENT	TFR
Afghanistan	Asia	5.27
Timor-Leste	Asia	5.30
COUNTRY	**CONTINENT**	**TFR**
Australia	Oceania	1.92

Reorder by median



Variable width boxplots

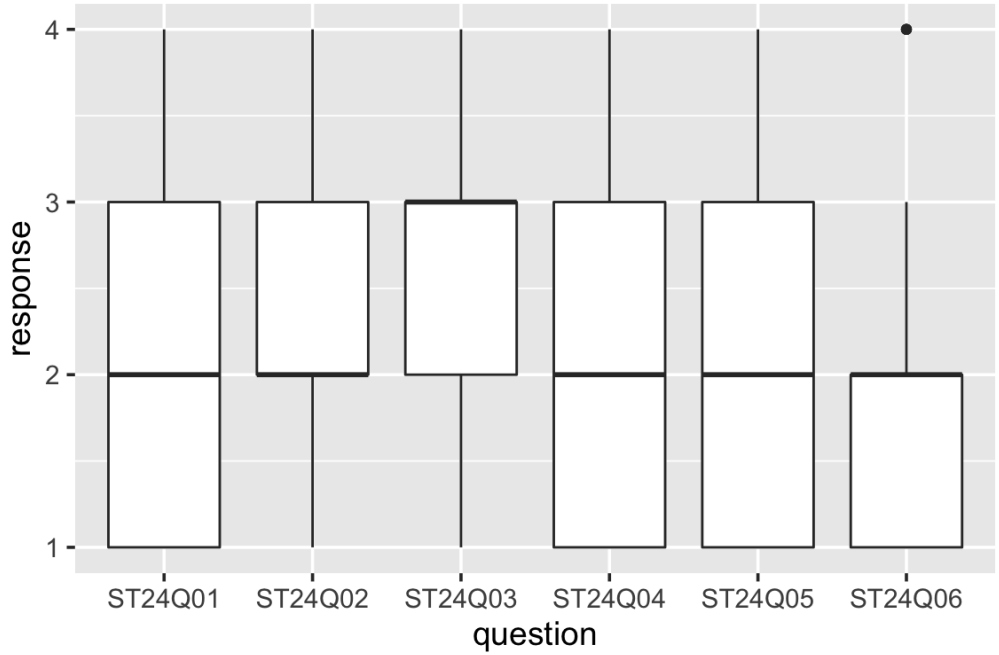
```
+ geom_boxplot(varwidth = TRUE)
```



width of boxes is proportional to $\sqrt{n}$

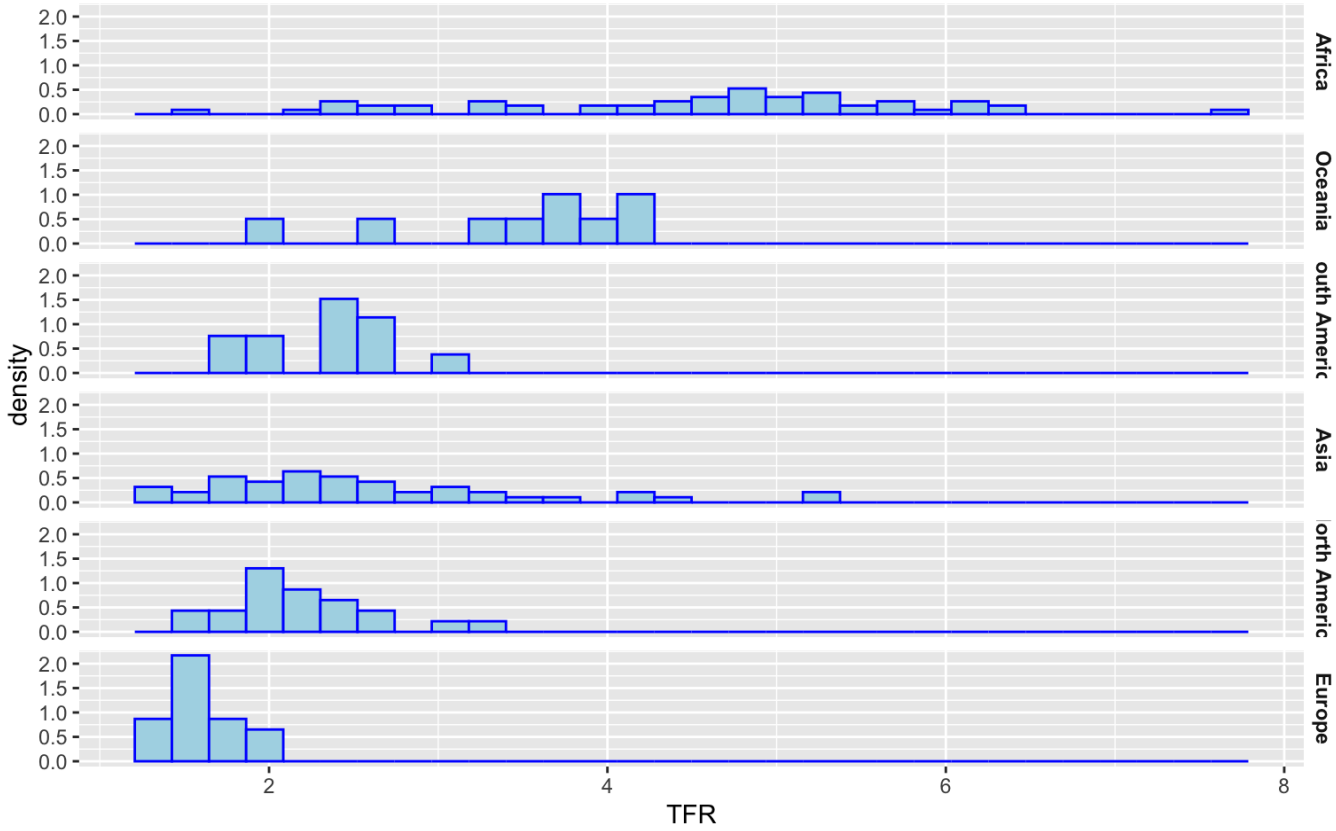
(Counts are 52, 9, 12, 43, 21, 42)

Not for discrete data

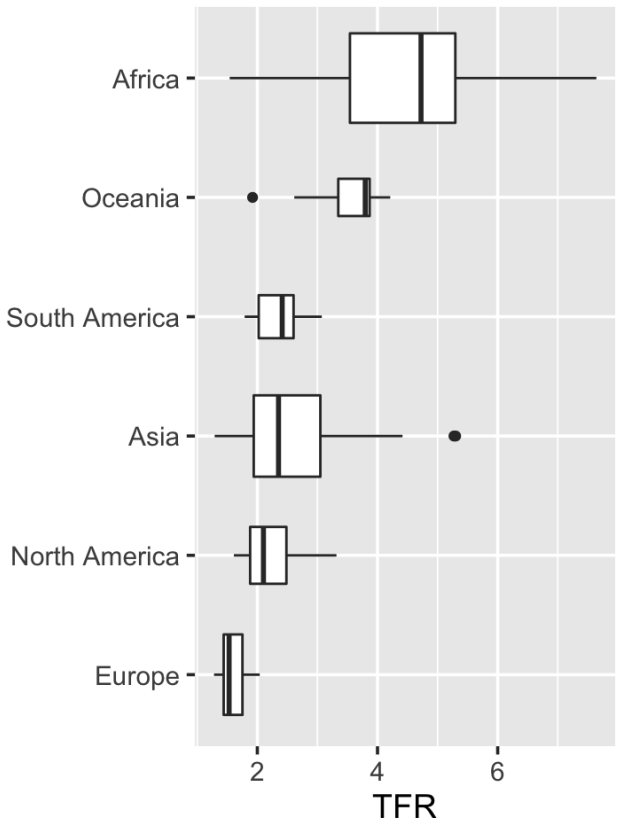


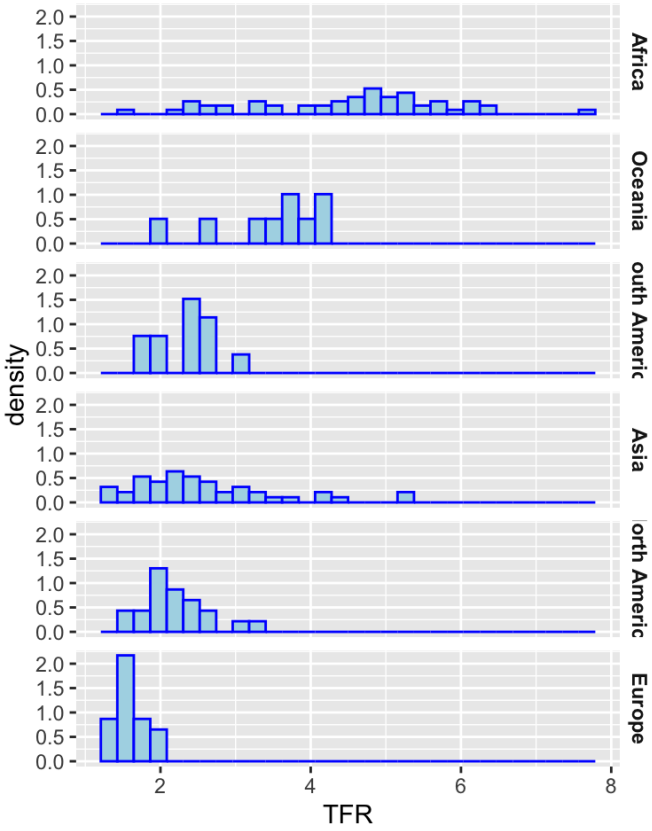
Source: R likert::pisaitems dataset

Multiple density histograms, ordered by median

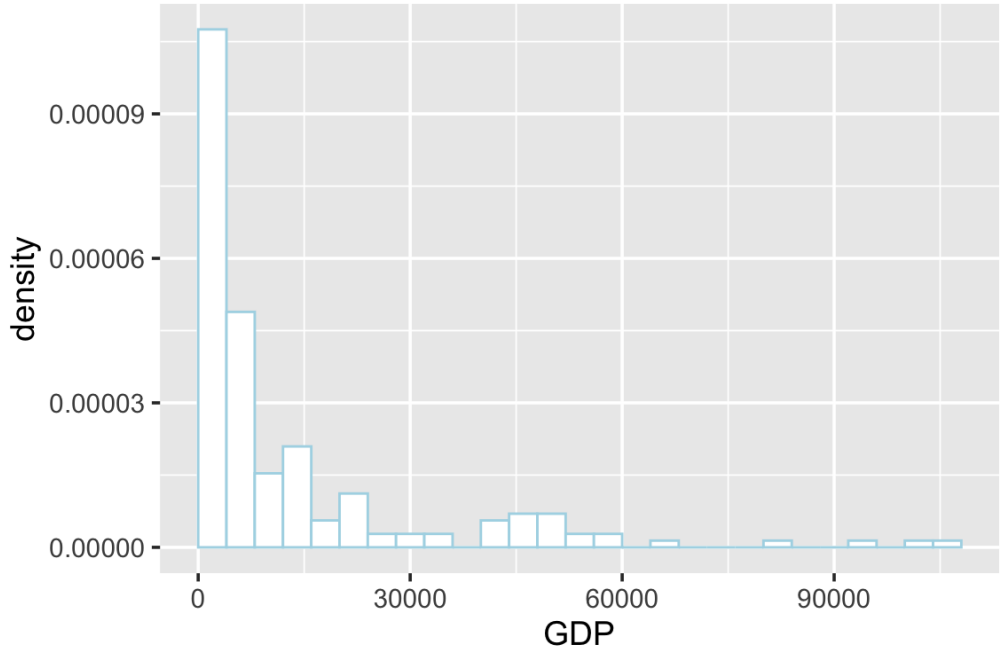


Boxplots vs. histograms



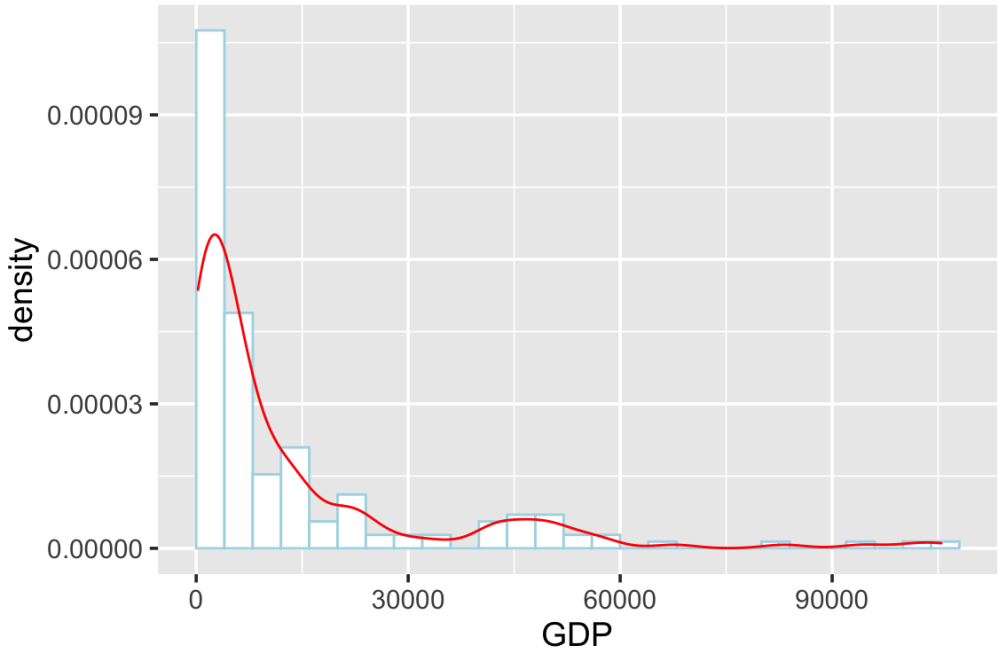


Density histogram



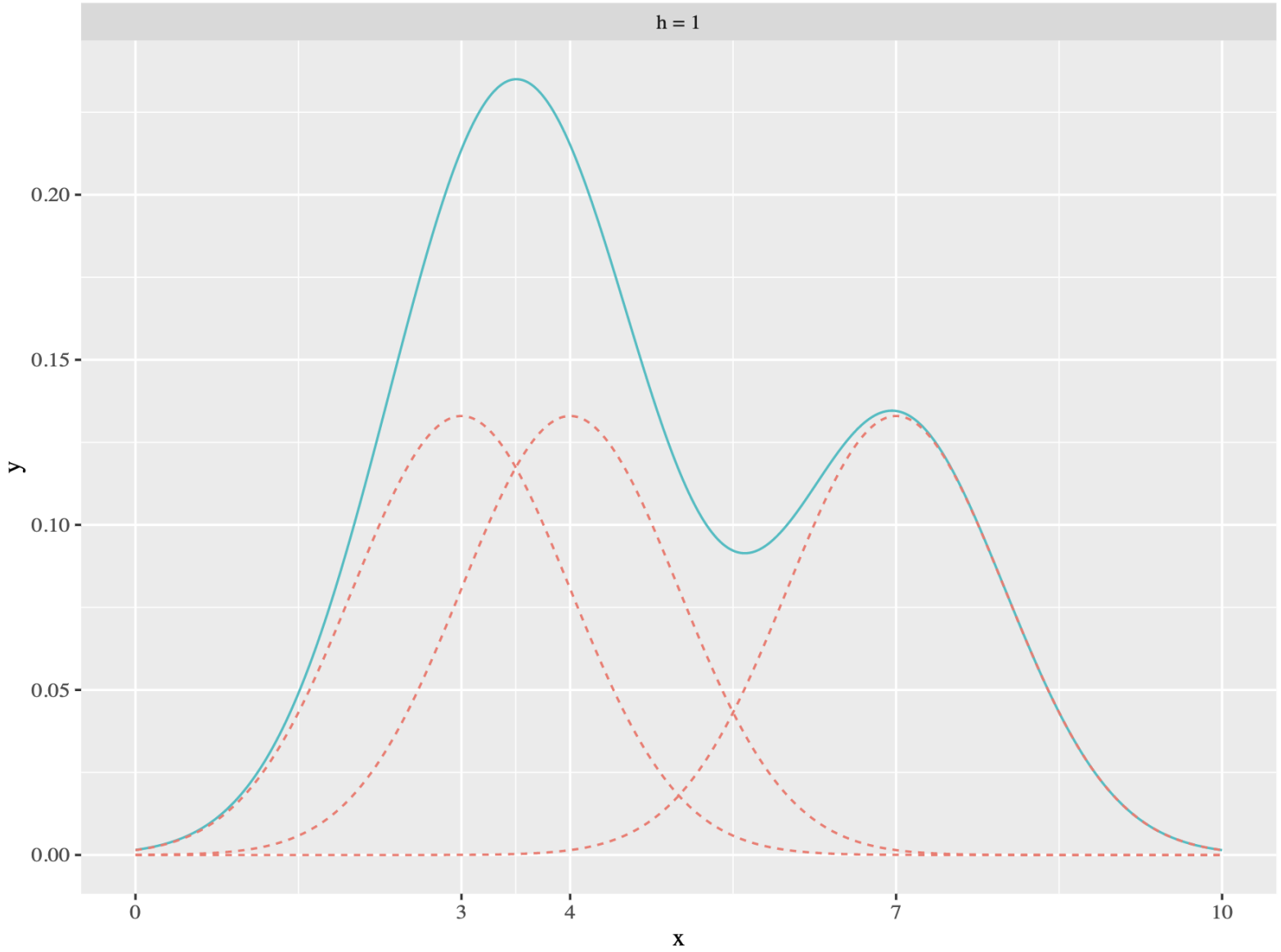
Density curve

```
+ geom_density()
```

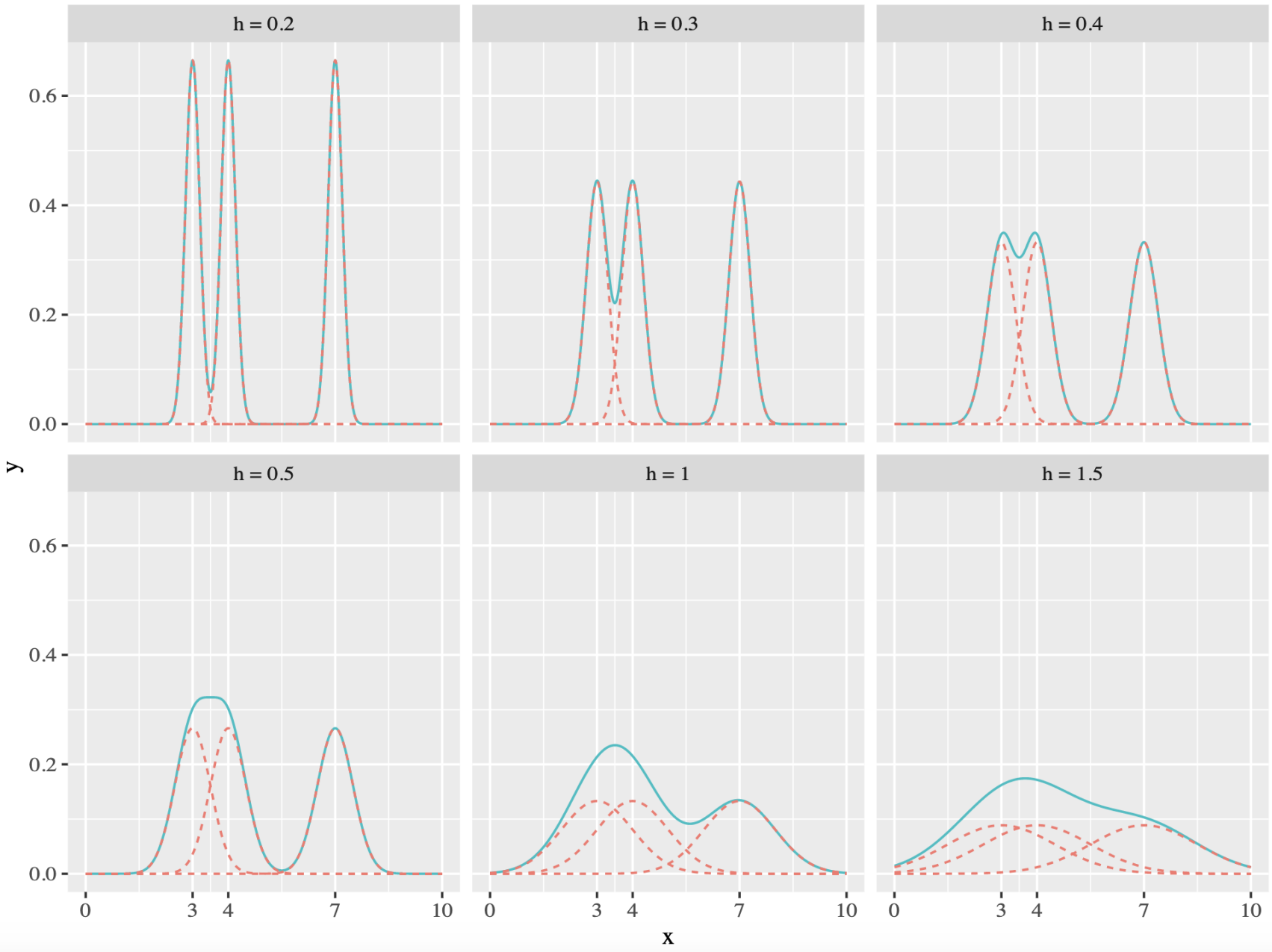


Density curve

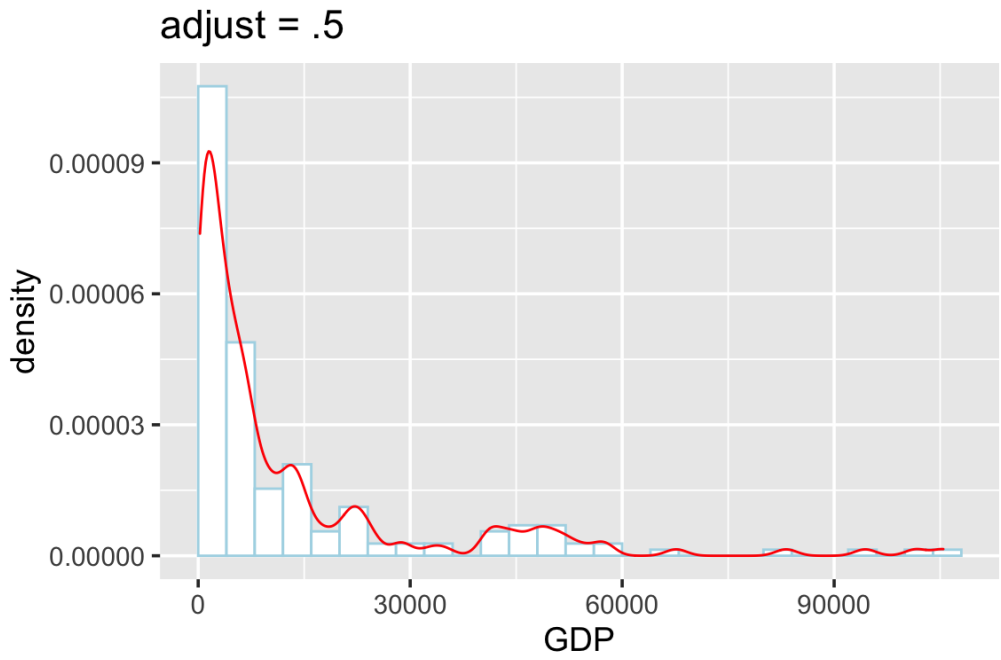
- assuming a normal gaussian distribution-



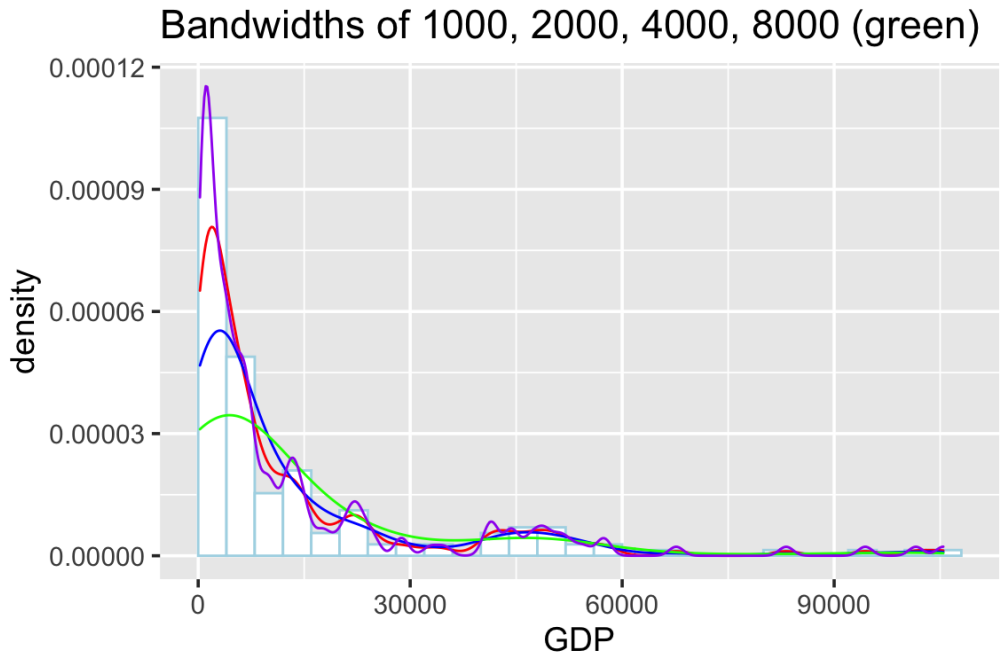
Density curve



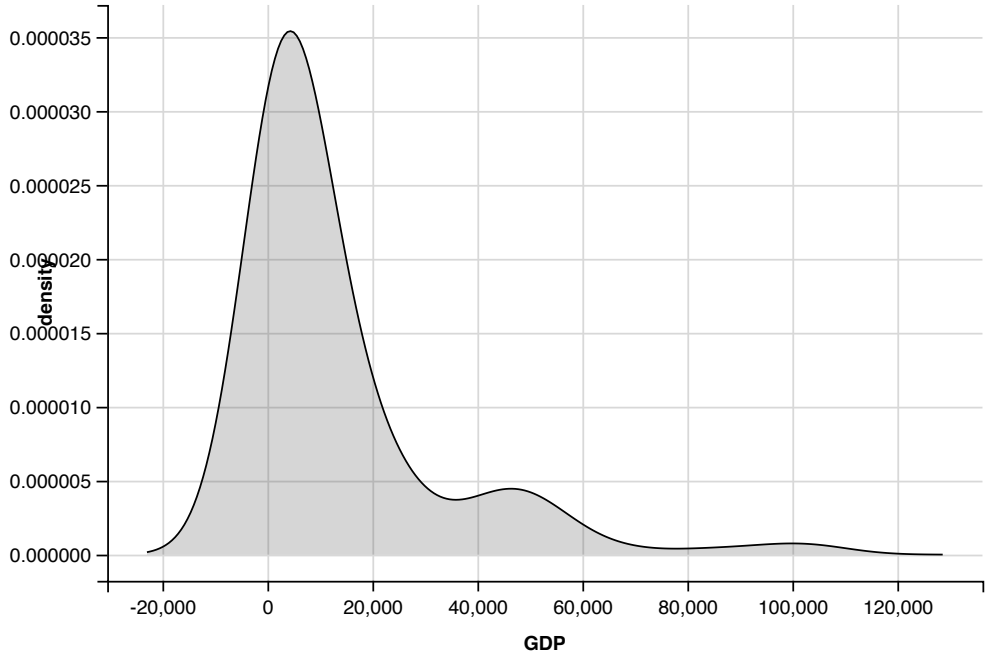
Density curve



Density curve: varying smoothing bandwidths



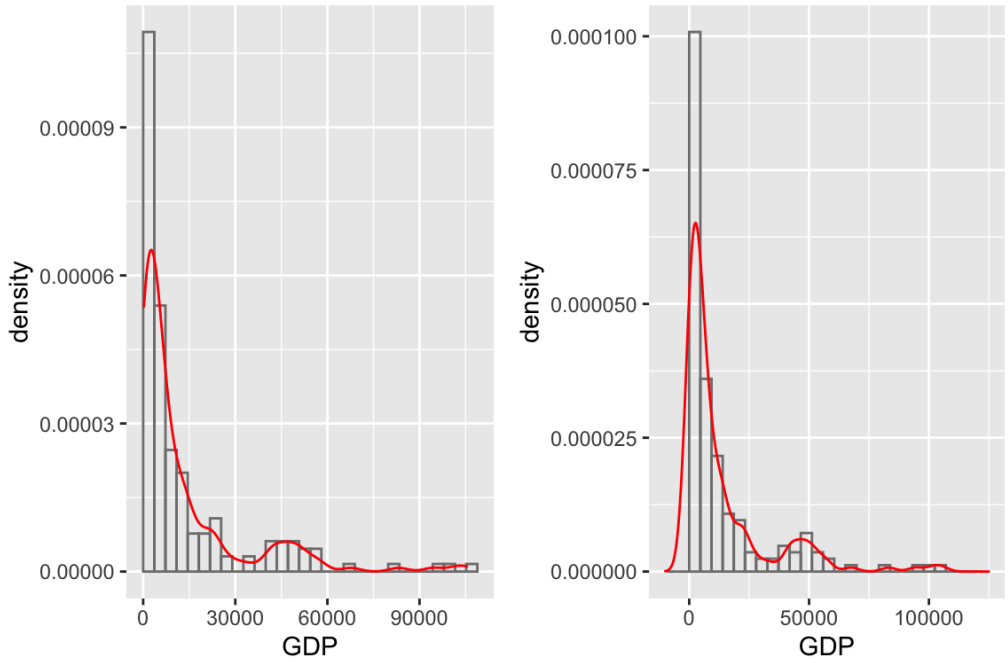
Density curve: varying smoothing bandwidths (`ggvis`)



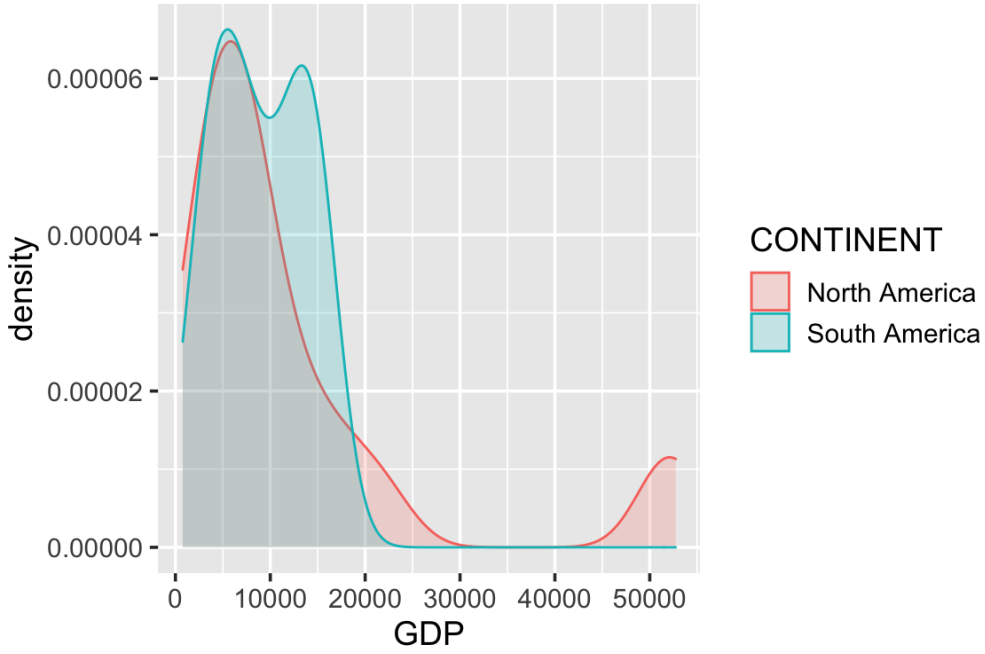
See also: http://ggvis.rstudio.com/0.1/quick-examples.html#histograms

x-axis limits

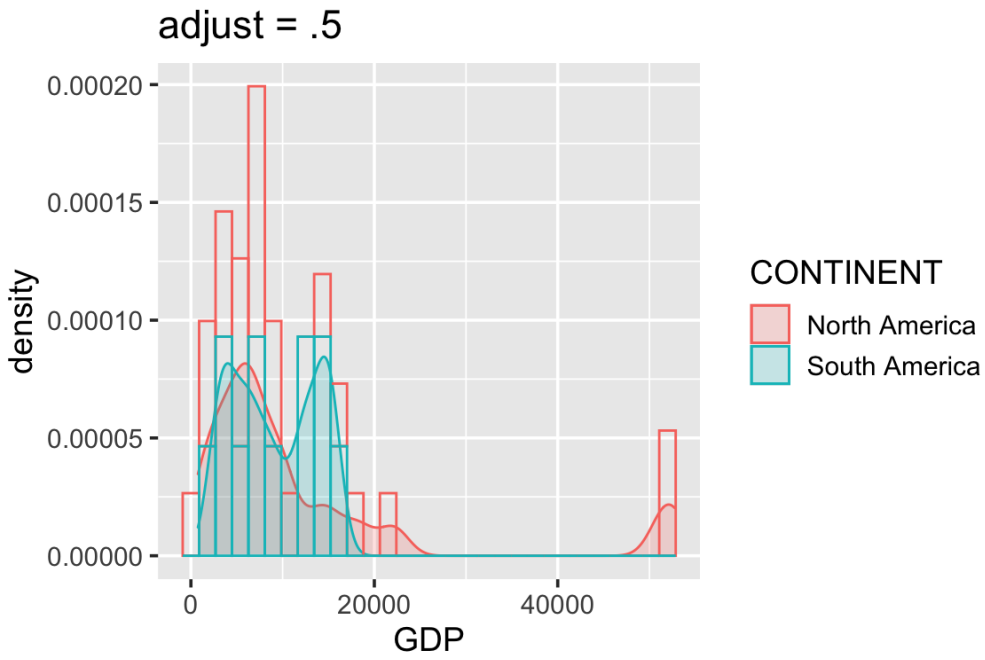
```
g1 <- ggplot(world, aes(GDP, y = ..density..)) +
  geom_histogram(color = "grey50", fill = NA, boundary = 0) +
  geom_density(color = "red")
g2 <- ggplot(world, aes(GDP, y = ..density..)) +
  geom_histogram(color = "grey50", fill = NA, boundary = 0) +
  geom_density(color = "red") +
  scale_x_continuous(limits = c(-10000, 125000))
gridExtra::grid.arrange(g1, g2, nrow = 1)
```



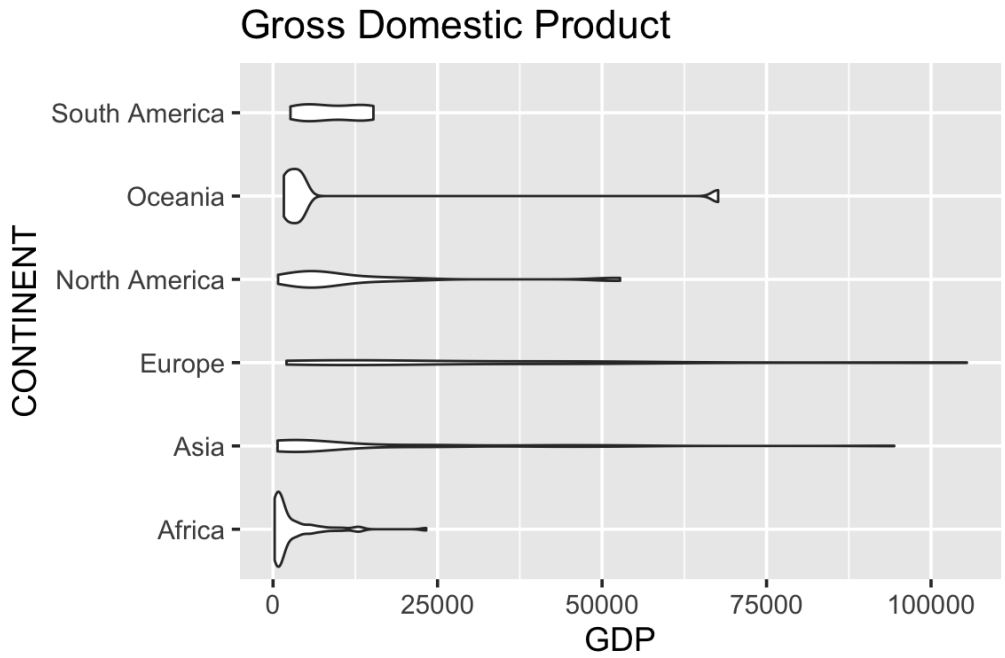
Density curves



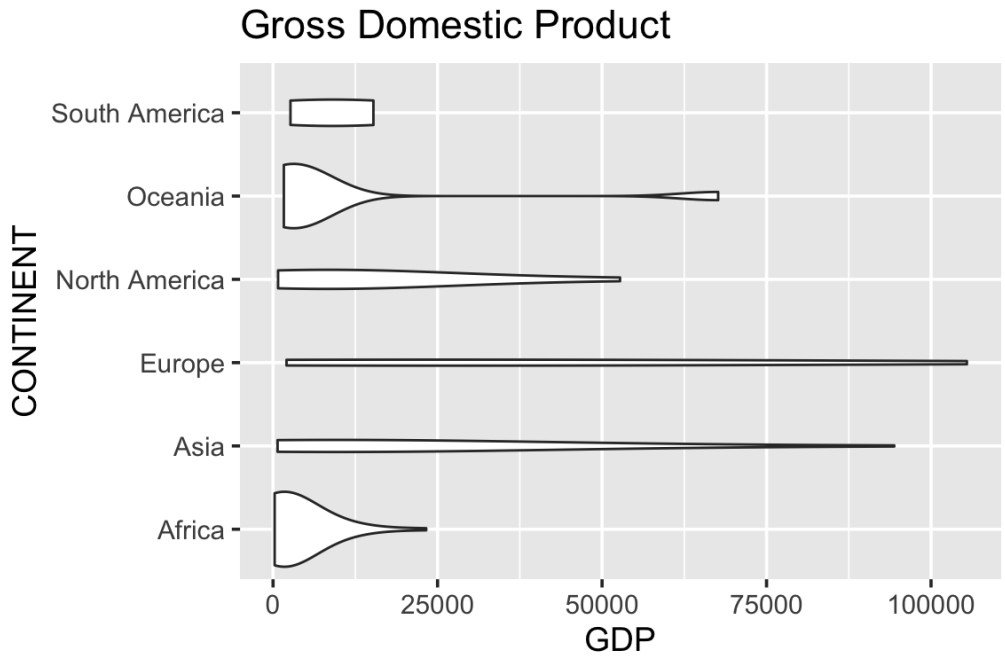
Density curves



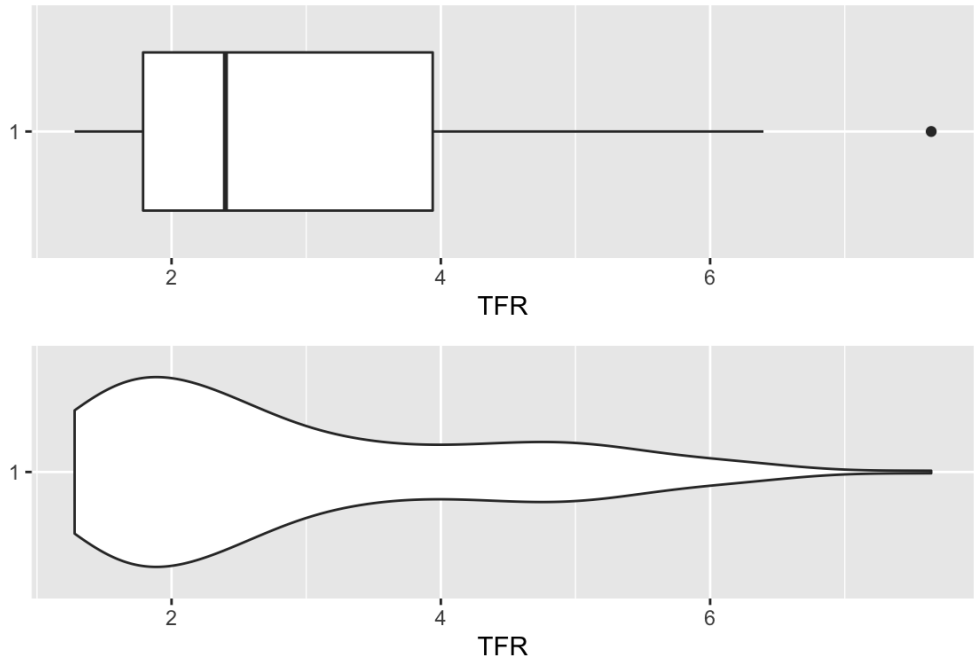
Violin plots



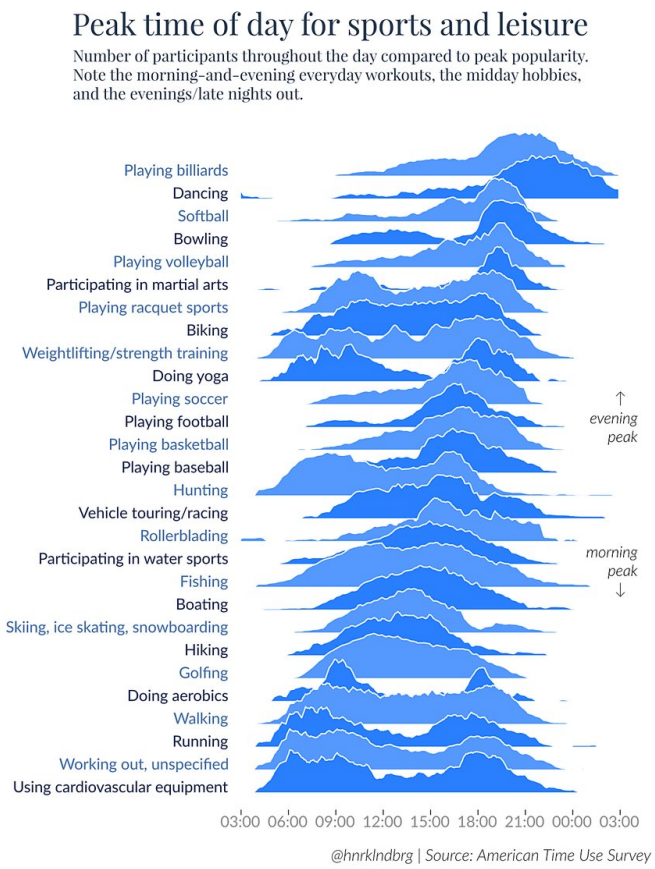
Violin plots, change bandwidth



Boxplot vs. violin plot



Ridgeline plot - basically a density curve with a little detail (sometimes interesting and useful and sometimes its not)



Source: https://eagereyes.org/blog/2017/joy-plots

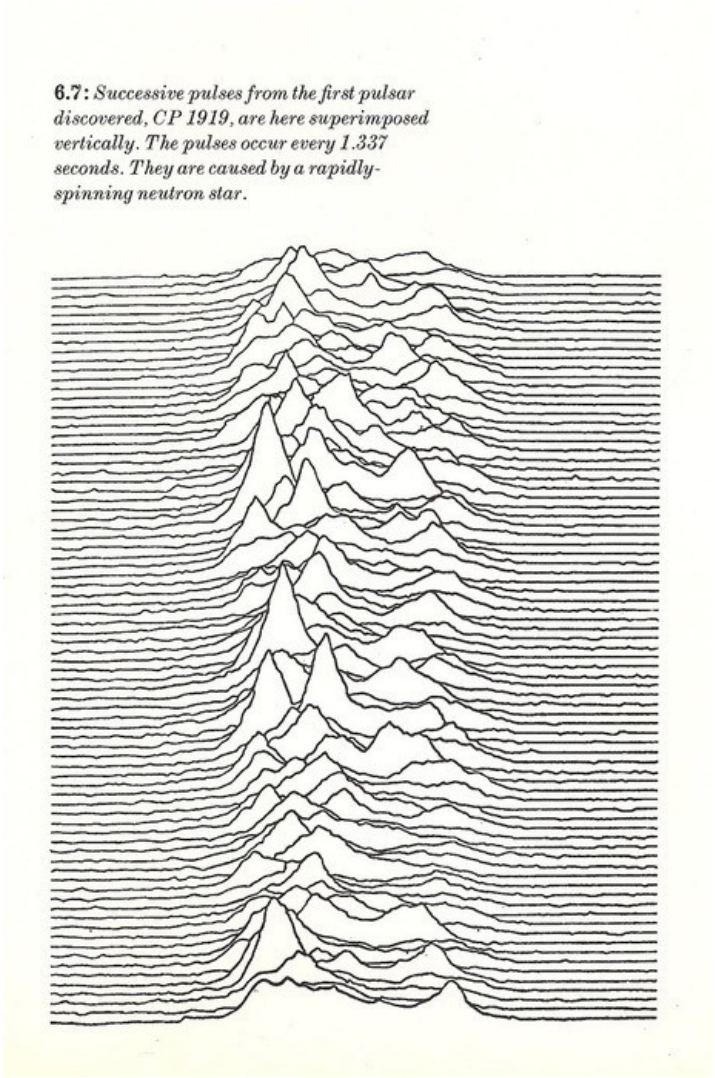
Additional resources:

http://blog.revolutionanalytics.com/2017/07/joyplots.html

https://blogs.scientificamerican.com/sa-visual/pop-culture-pulsar-origin-story-of-joy-division-s-unknown-pleasures-album-cover-video/

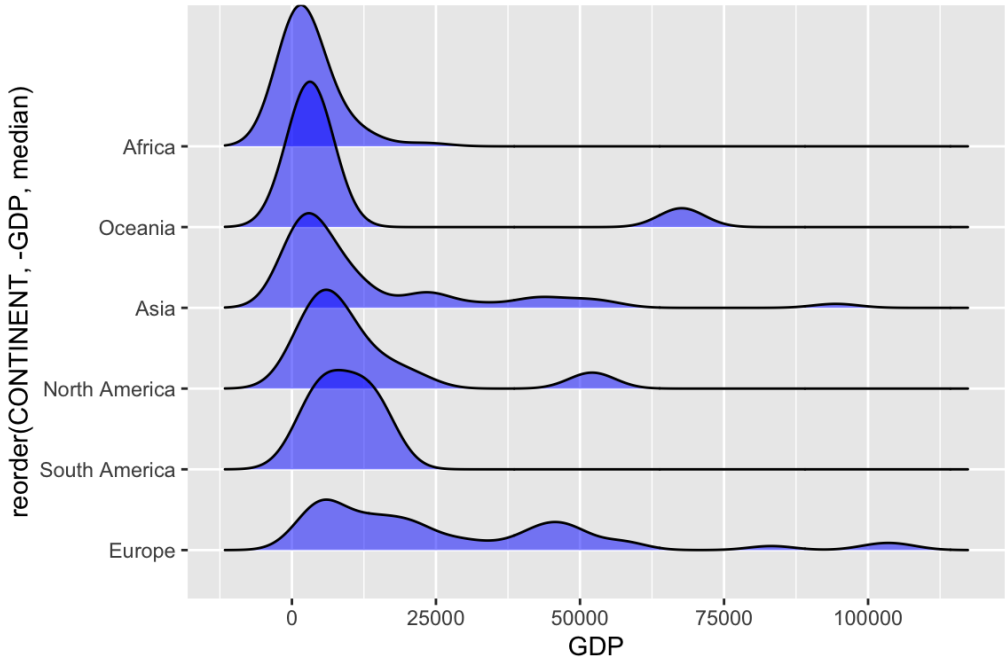
Ridgeline plot inspiration

Jocelyn Bell discovers first radio pulsars, 1967

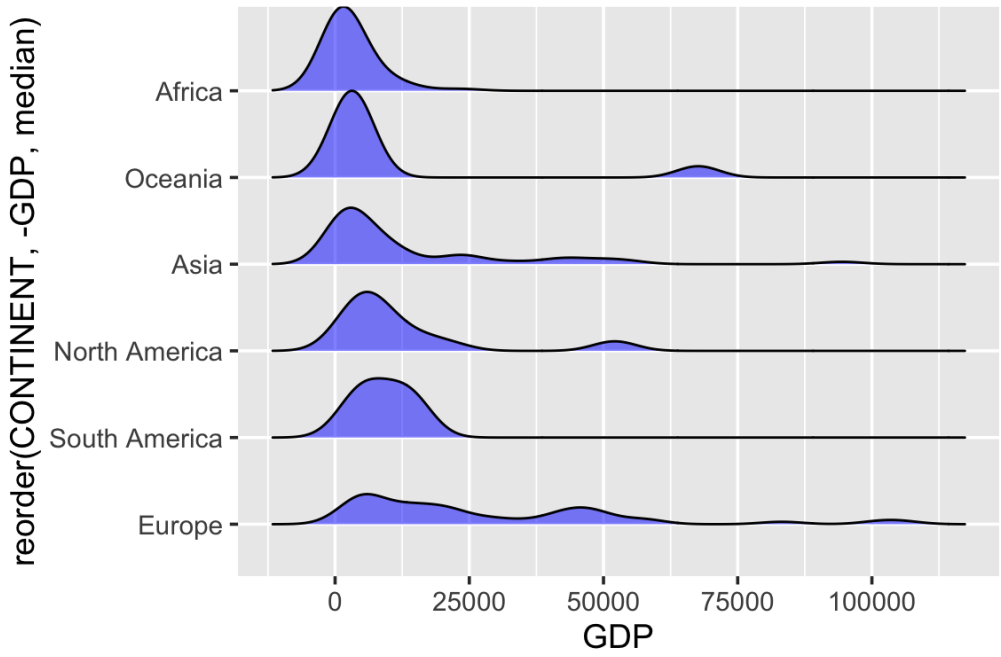


6.7: *Successive pulses from the first pulsar discovered, CP 1919, are here superimposed vertically. The pulses occur every 1.337 seconds. They are caused by a rapidly-spinning neutron star.*

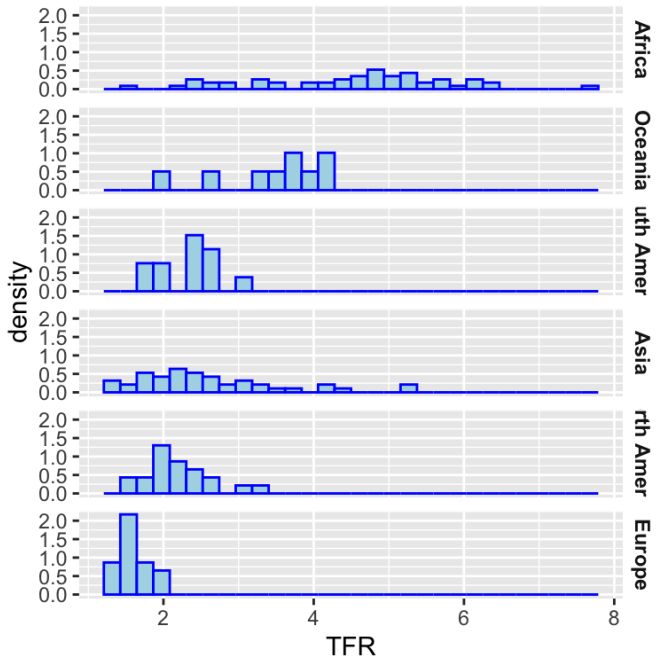
Ridgeline plot

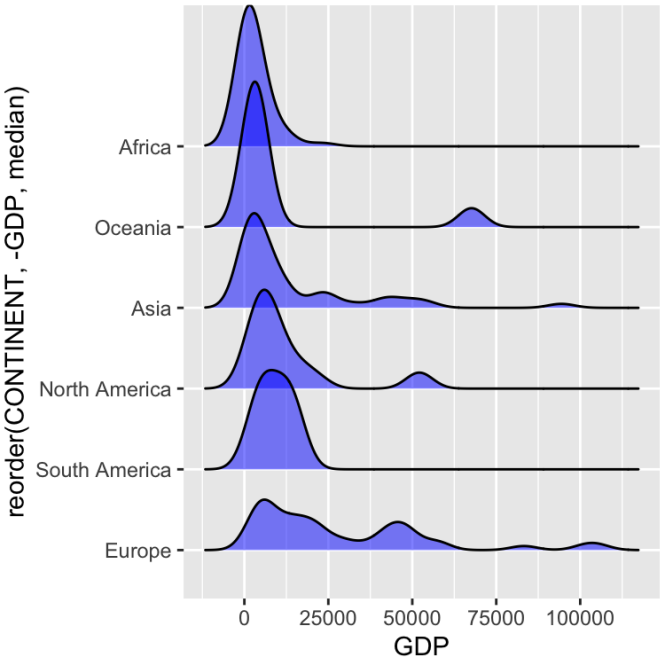


Ridgeline plot, change scale



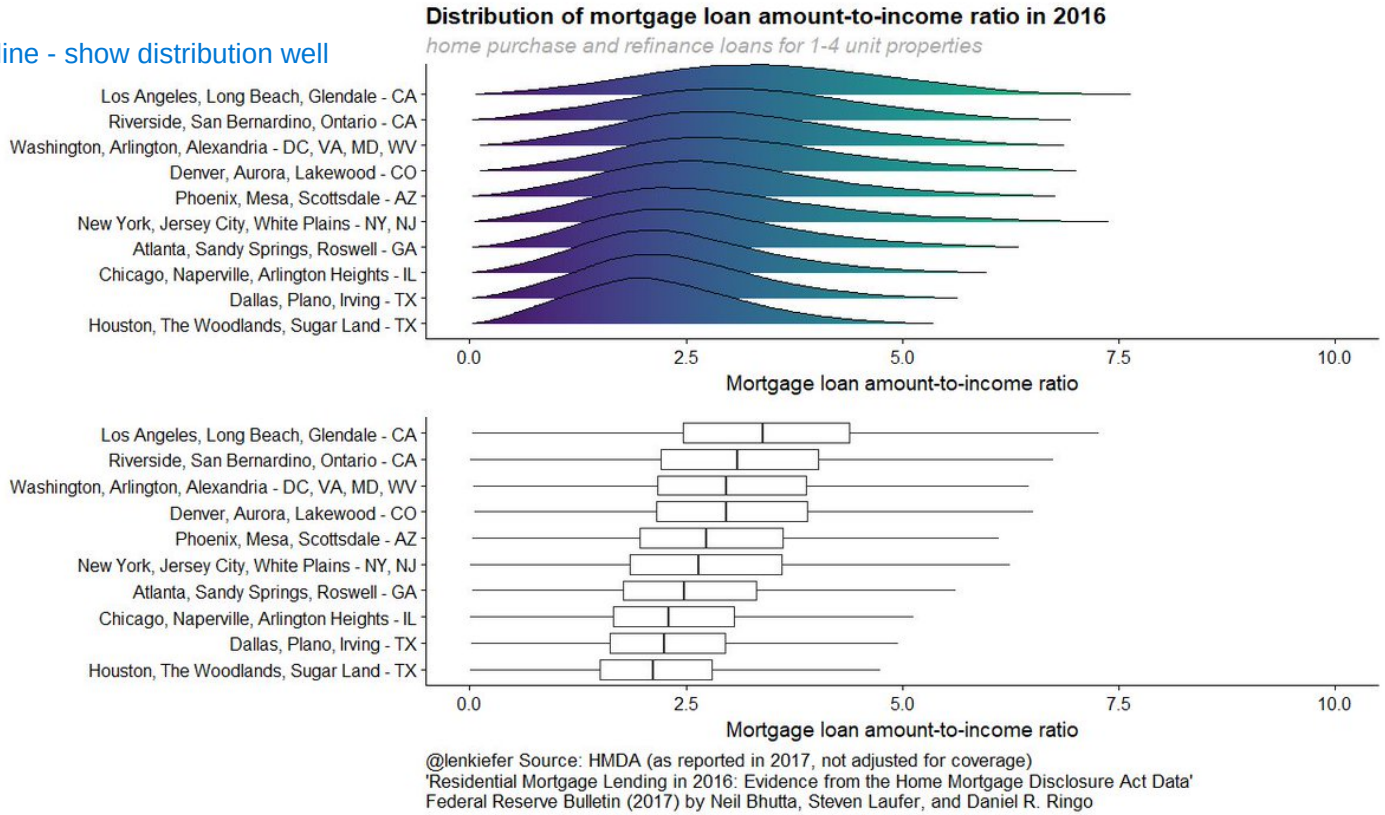
Histogram vs. ridgeline





Ridgeline vs. boxplot

ridgeline - show distribution well



Source: https://twitter.com/lenkiefer/status/916823350726610946

`ggridges` package

CRAN https://CRAN.R-project.org/package=ggridges

Github https://github.com/clauswilke/ggridges

Package vignette(s) https://cran.r-project.org/web/packages/ggridges/vignettes/introduction.html

https://cran.r-project.org/web/packages/ggridges/vignettes/gallery.html

Package manual https://cran.r-project.org/web/packages/ggridges/ggridges.pdf