GR5702
Exploratory Data Analysis and Visualization

Prof. Joyce Robbins

Today's Agenda (1/26/ 17)

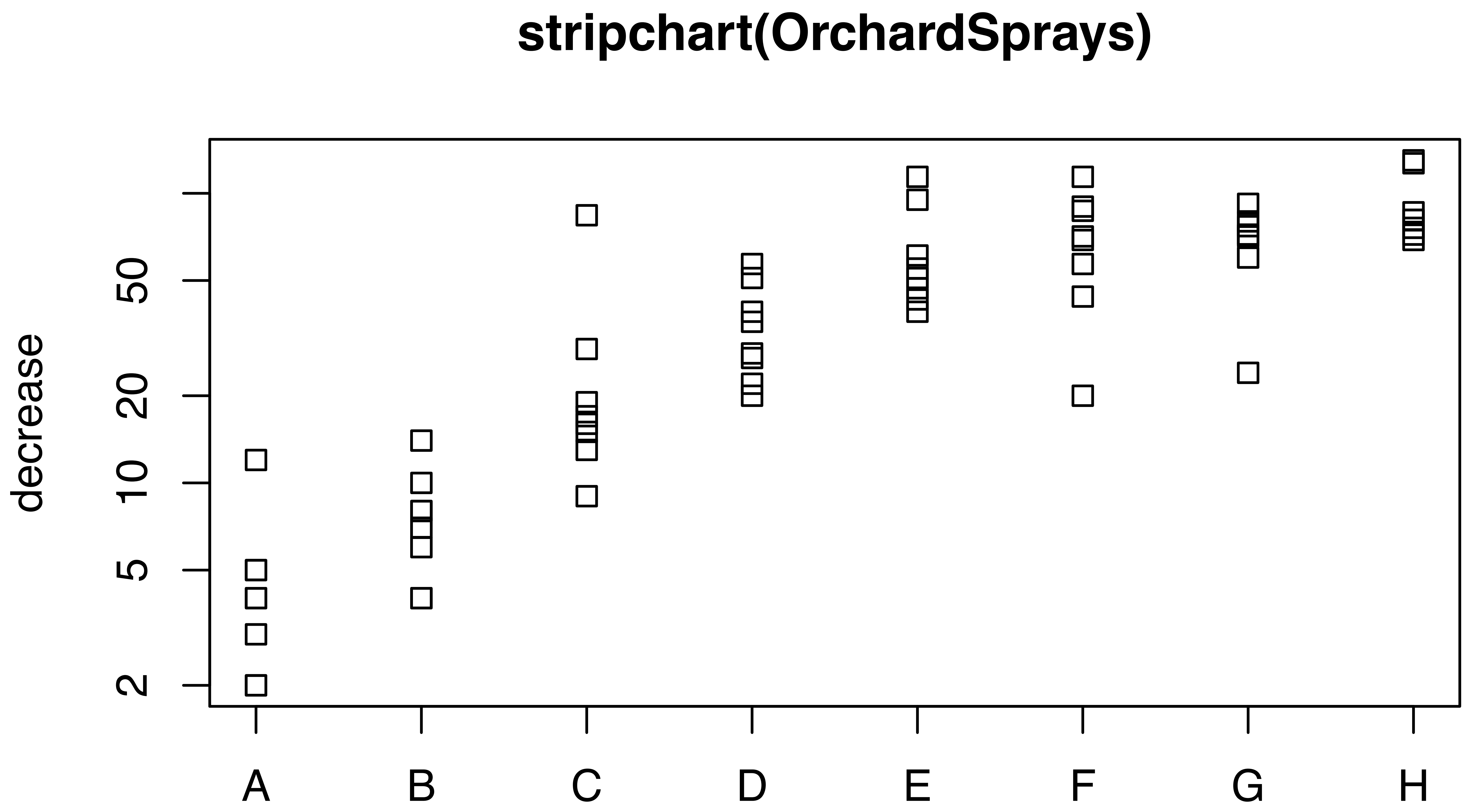
- Announcements
  - Final Project
  - Homework http://flowingdata.com
  - DataCamp http://datacamp.com
- Grammar of Graphics / ggplot2

One dimensional data

- individual ↔ distribution / group
- range
- summary statistics
- skewness
- frequency
- gaps
- outliers

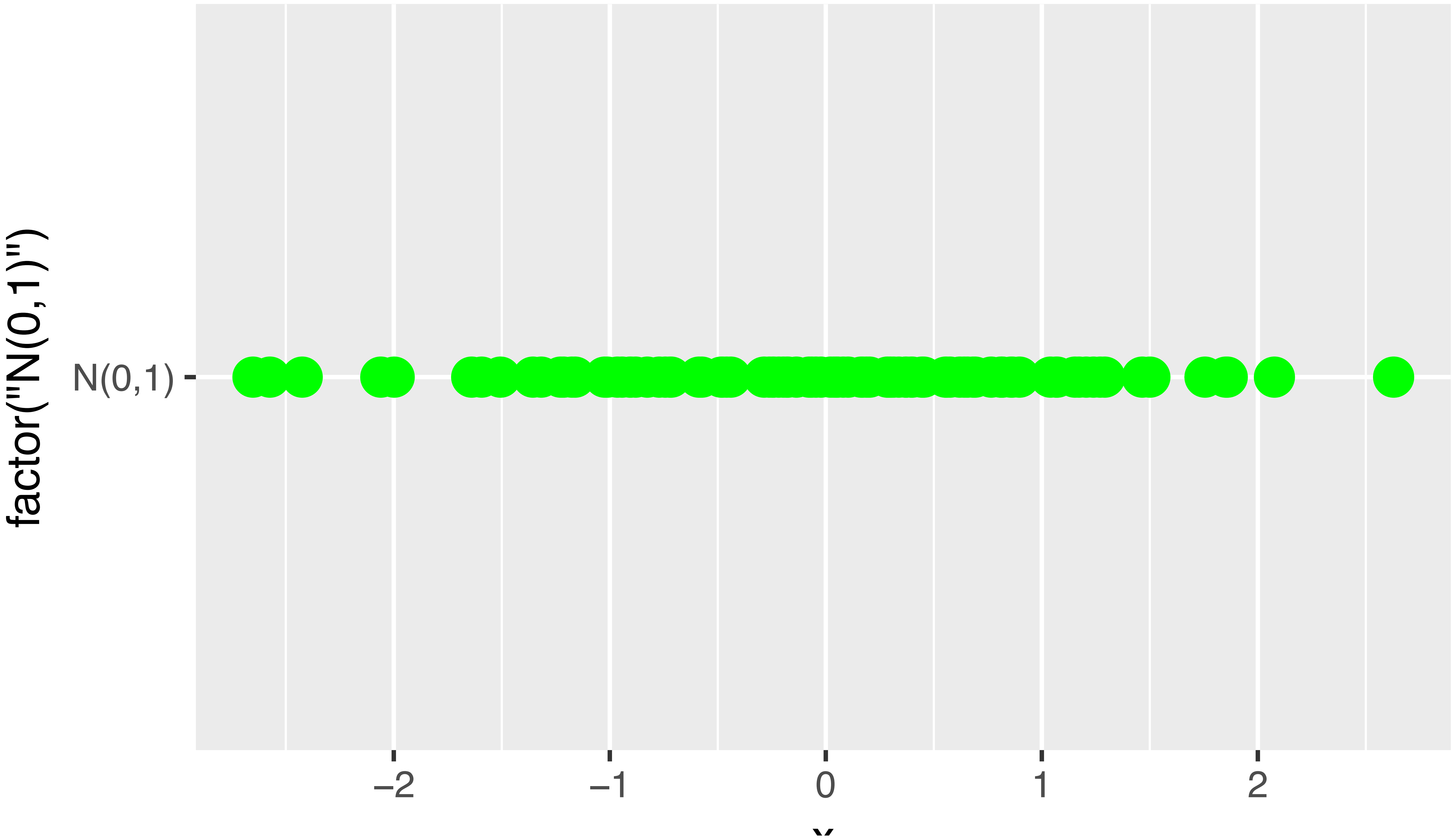
Strip charts

```
stripchart(decrease ~ treatment,
           main = "stripchart(OrchardSprays)",
           vertical = TRUE, log = "y", data = OrchardSprays)
```



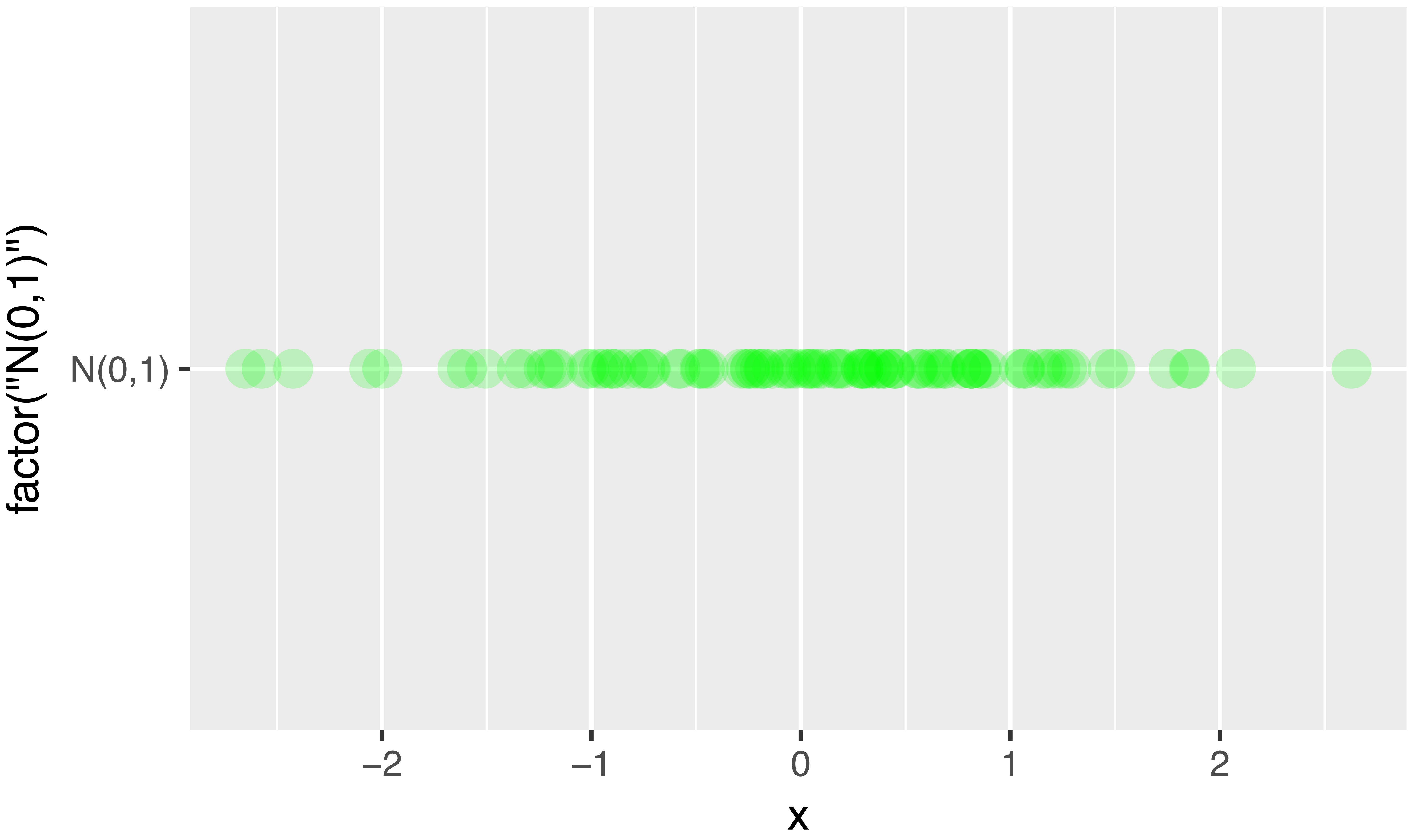
Strip plot (ggplot2)

```r
library(ggplot2)
x <- rnorm(100)
ggplot(data.frame(x), aes(x, y = factor("N(0,1)"))) +
    geom_point(size = 4, color = "green")
```



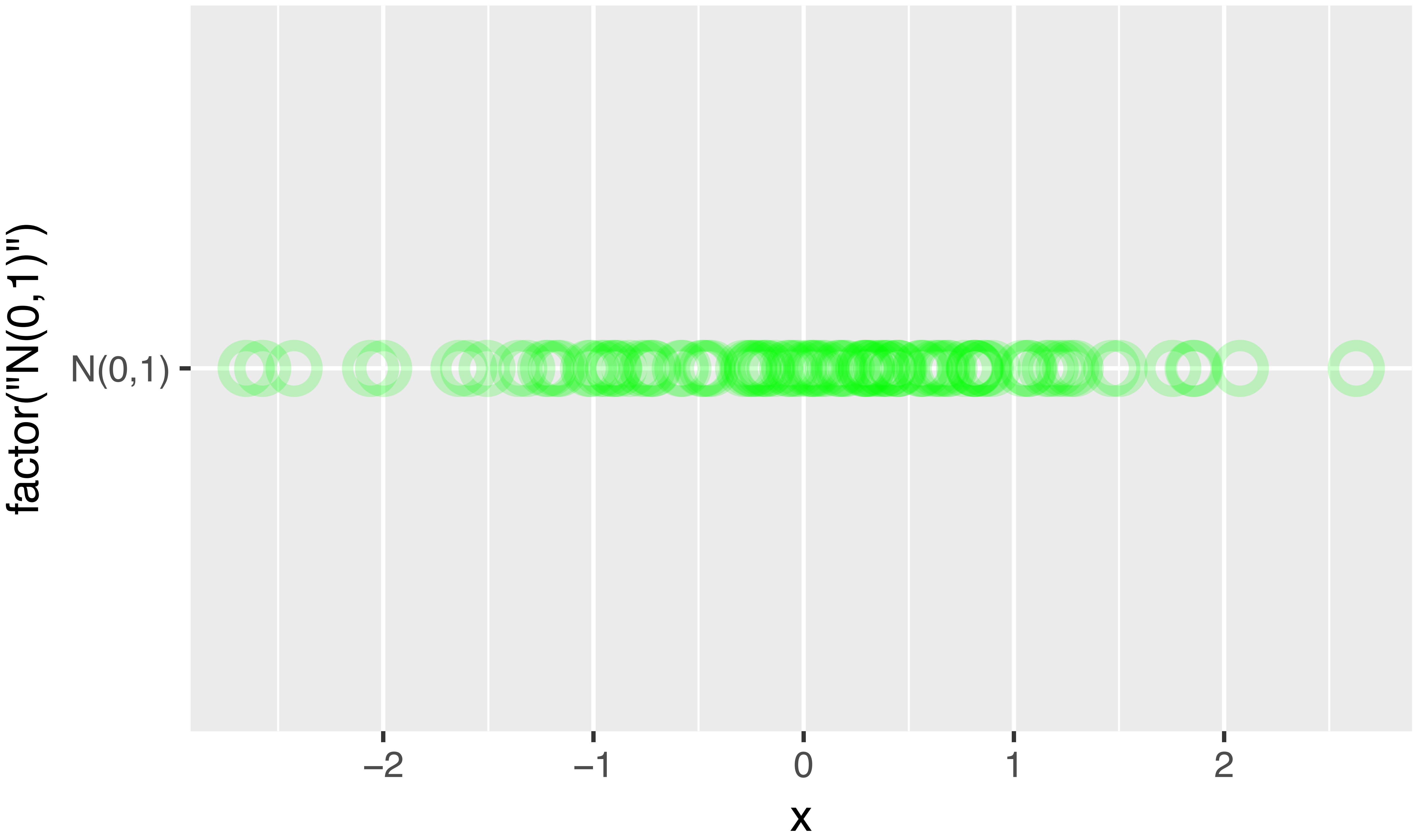
Strip plot (ggplot2) w/ alpha

```
ggplot(data.frame(x), aes(x, y = factor("N(0,1)"))) +
    geom_point(size = 4, color = "green", alpha = .2)
```



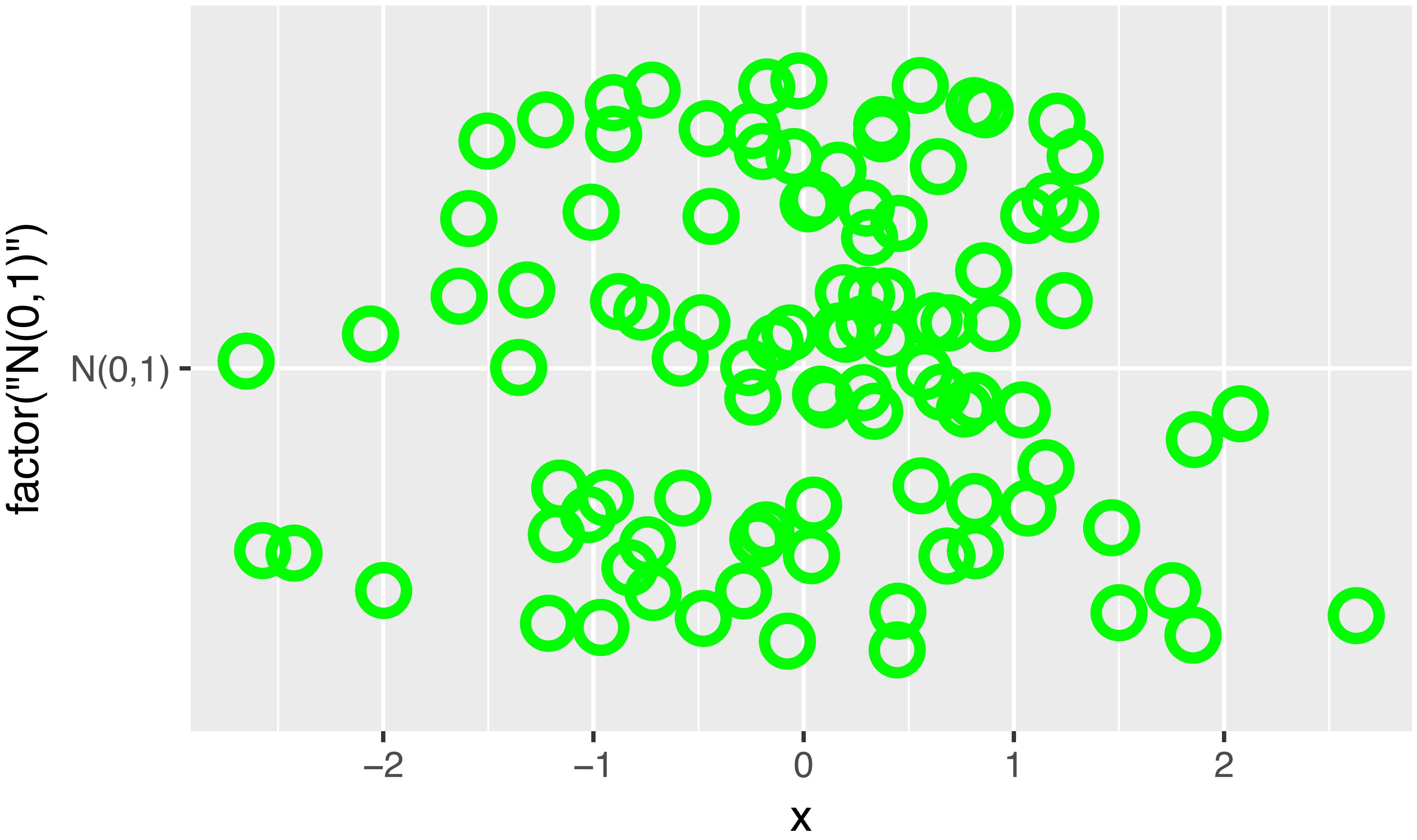
Strip plot (ggplot2) w/ alpha, shape, stroke

```
ggplot(data.frame(x), aes(x, y = factor("N(0,1)"))) +
    geom_point(size = 4, color = "green", alpha = .2, shap
                stroke = 2)
```



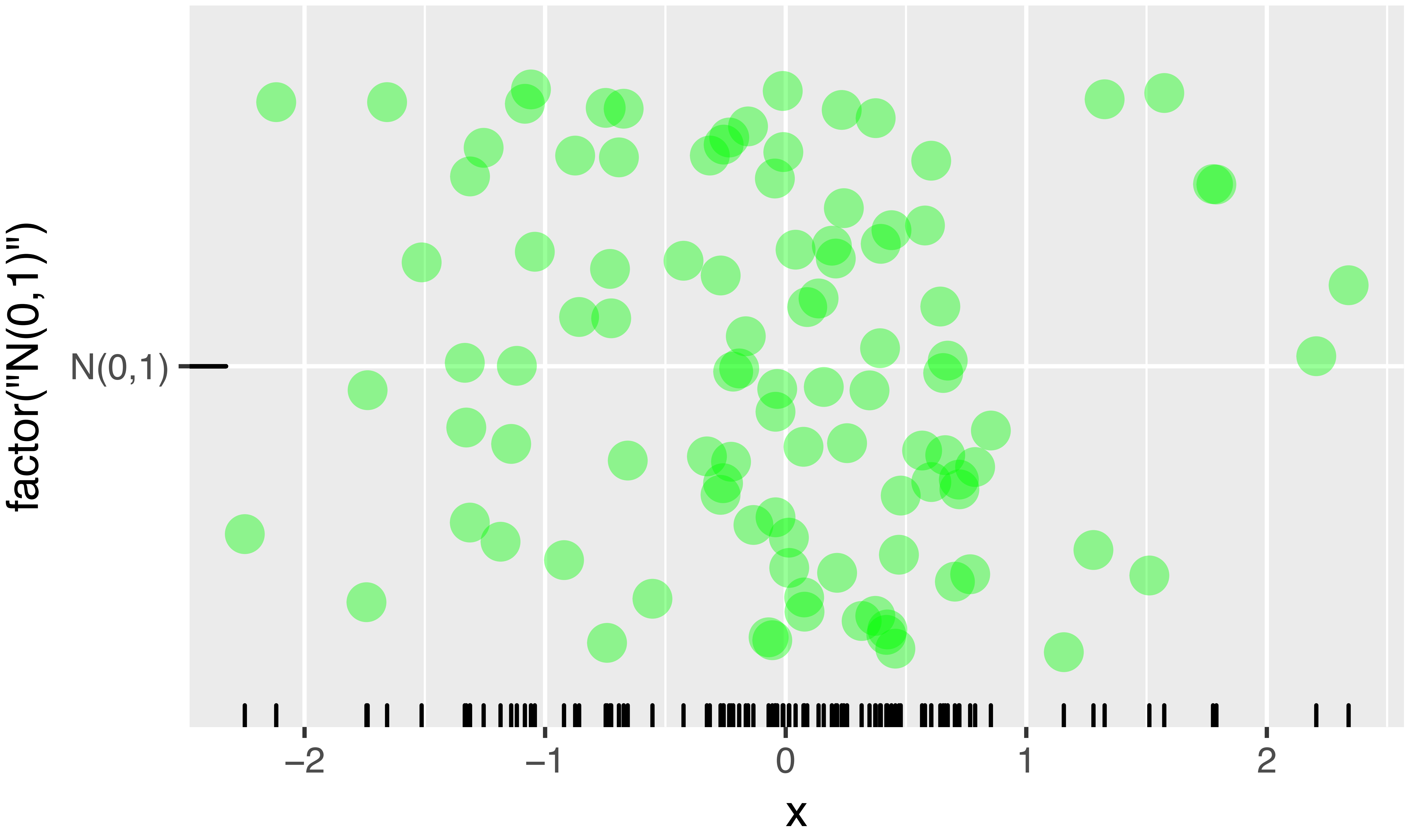
Strip plot (ggplot2) w/ jitter

```r
ggplot(data.frame(x), aes(x, y = factor("N(0,1)"))) +
    geom_point(size = 4, color = "green", shape = 1,
               stroke = 2, position = "jitter")
```



Strip plot (ggplot2) w/ jitter, alpha, fill, rug

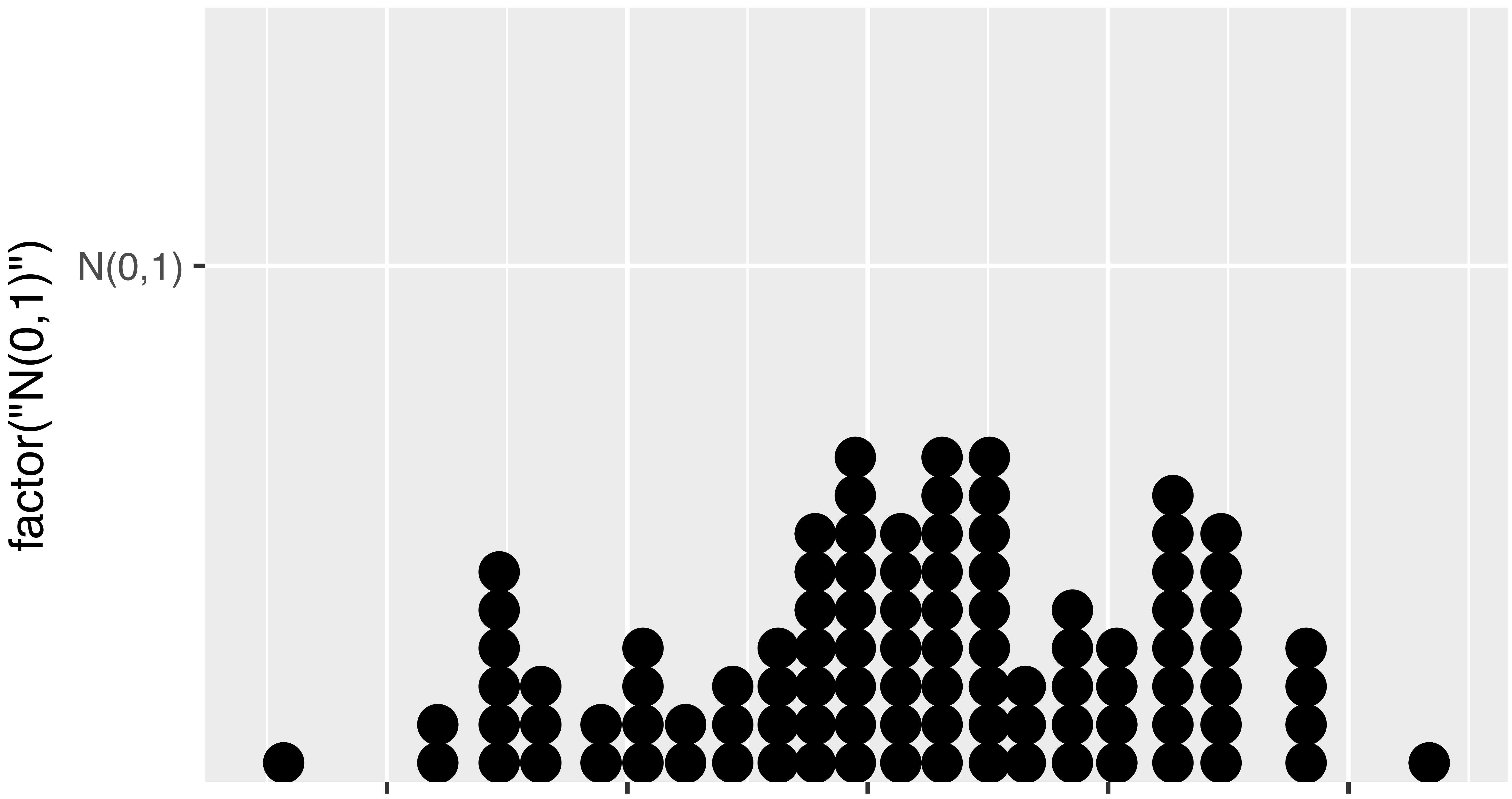
```
ggplot(data.frame(x), aes(x, y = factor("N(0,1)"))) +
    geom_point(size = 4, color = "green", alpha = .4,
               position = "jitter") + geom_rug()
```



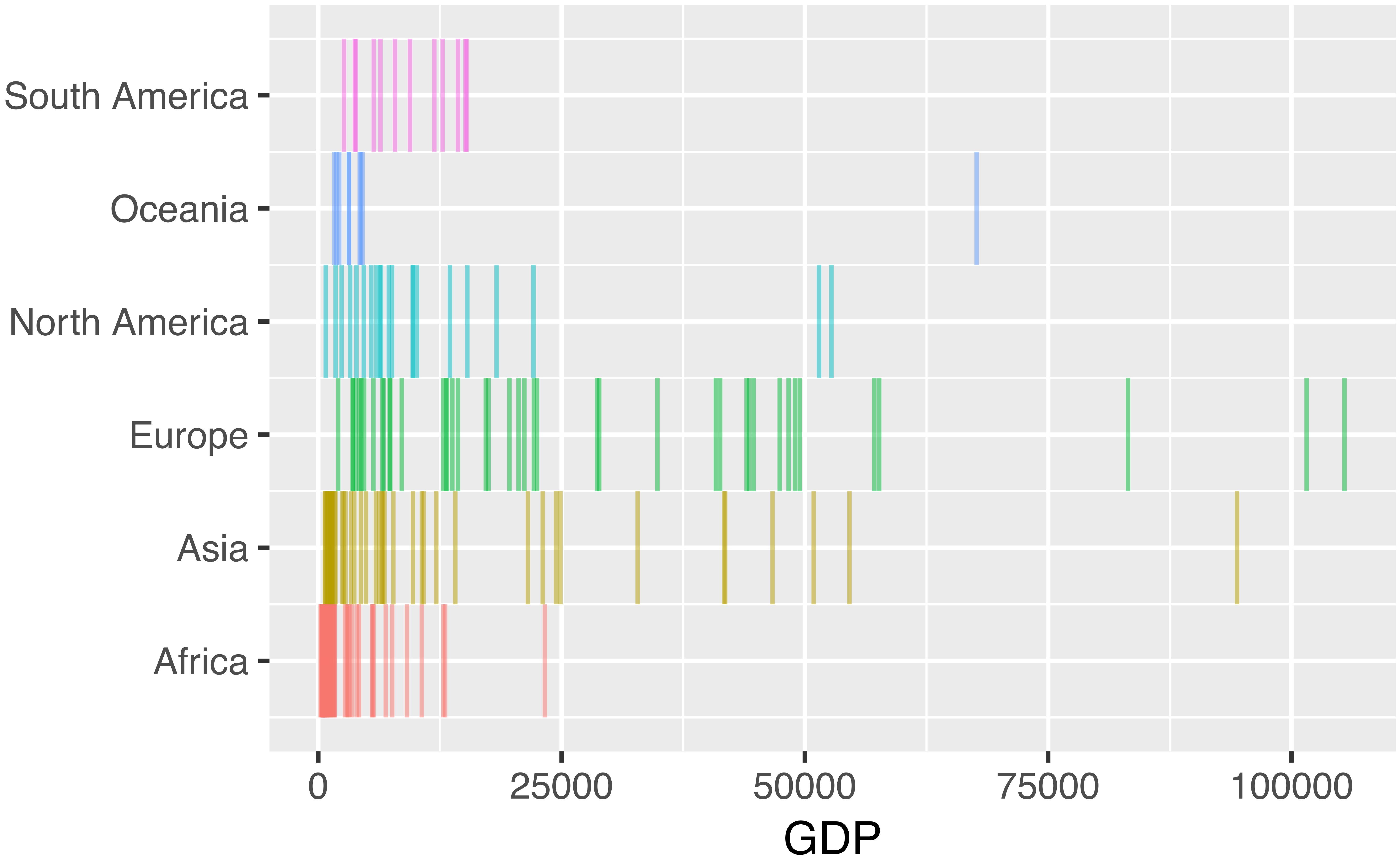
Dot plot

```
x <- rnorm(100)
ggplot(data.frame(x), aes(x, y = factor("N(0,1)"))) +
    geom_dotplot()
```

```
## `stat_bindot()` using `bins = 30`. Pick better value wi
```



Woven rug plot

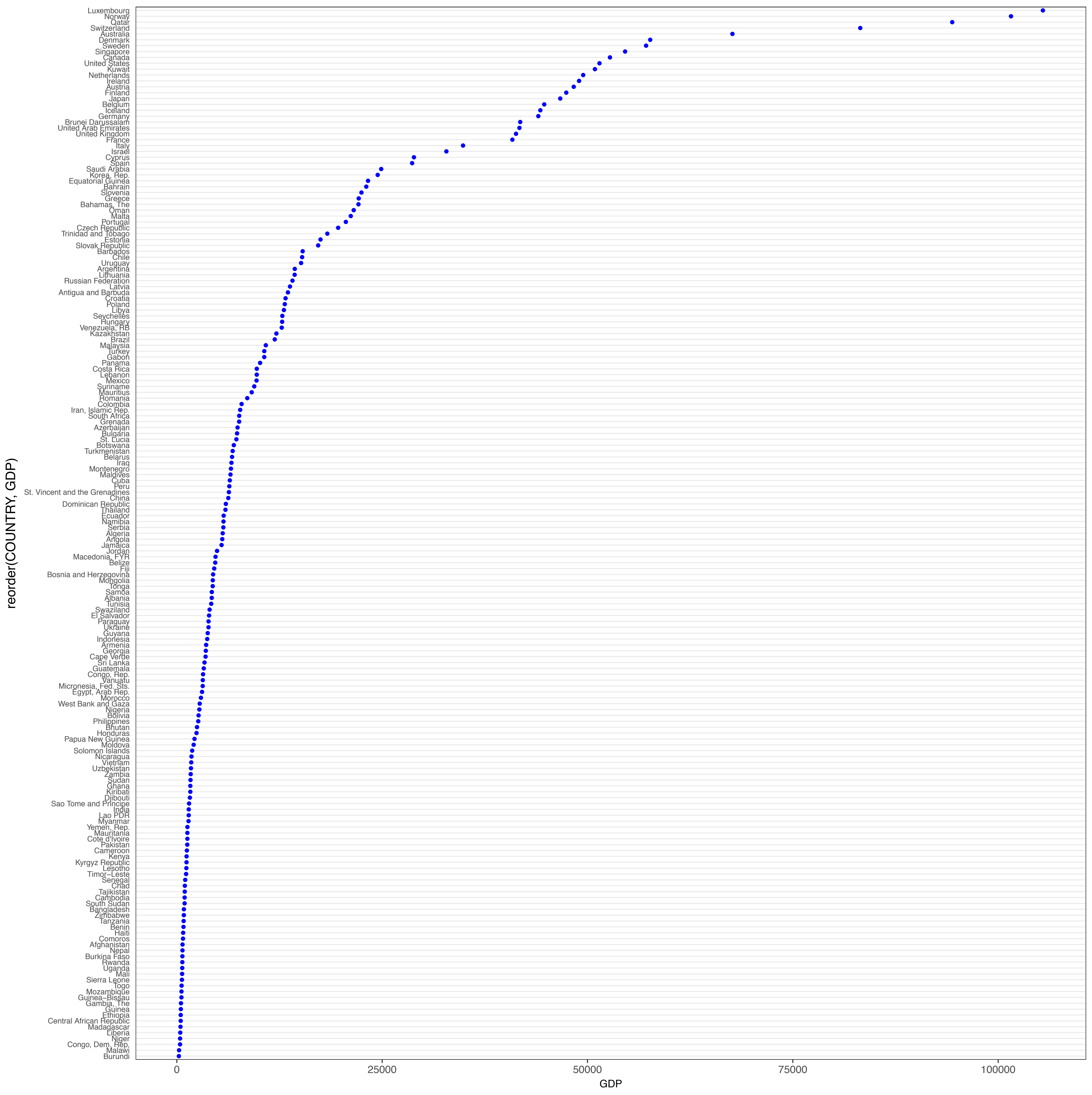


Woven rug code

```r
world <- read.csv("countries2012.csv", header = TRUE)

ggplot(world, aes(x = GDP, xend = GDP,
                  y = as.numeric(CONTINENT)-.5,
                yend = as.numeric(CONTINENT)+.5,
                color = CONTINENT)) +
    geom_segment(alpha = .5) +
    scale_y_continuous("", breaks = 1:6,
                       labels = levels(world$CONTINENT)) +
    guides(color = FALSE)
```

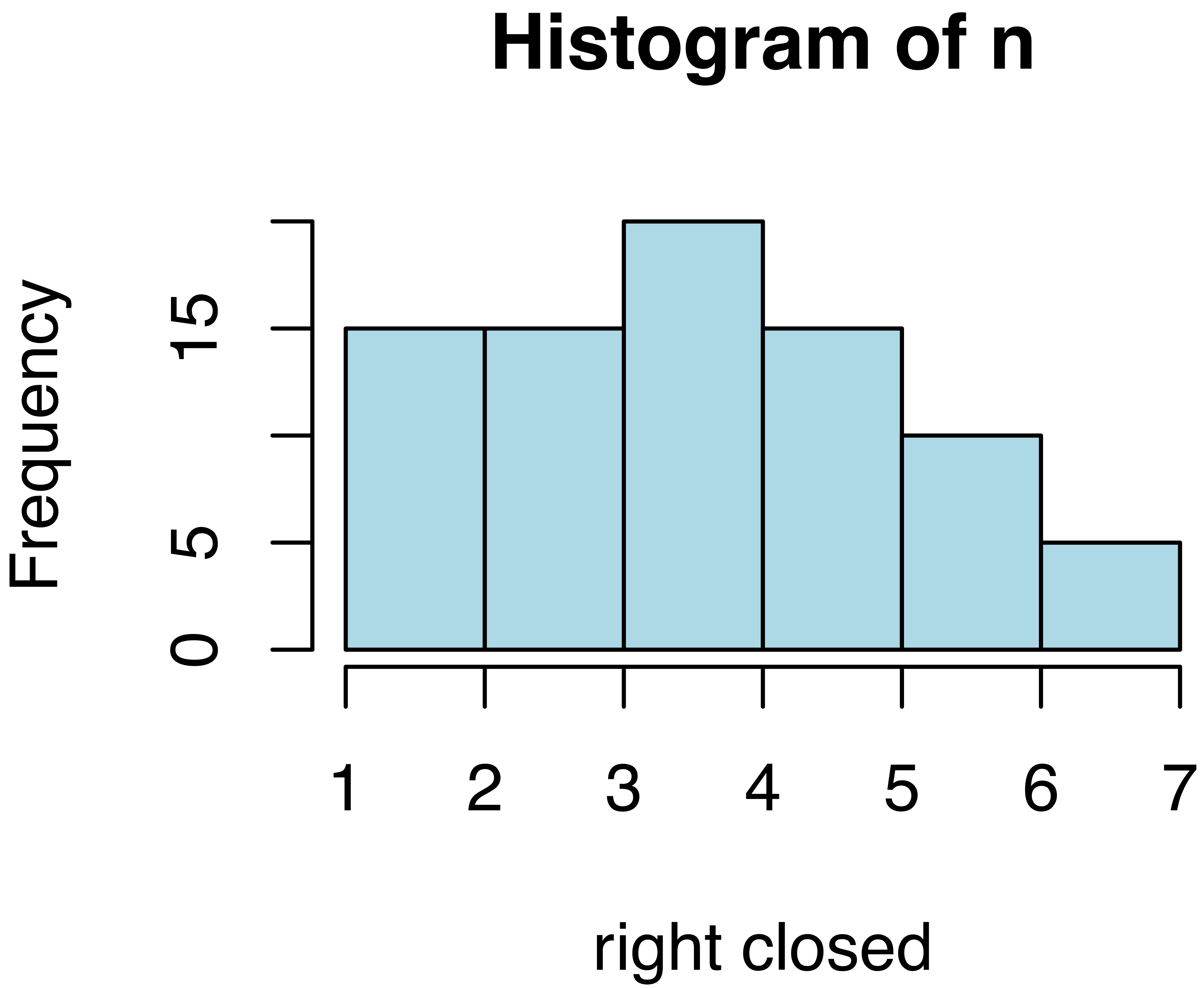
Cleveland Dot Plot



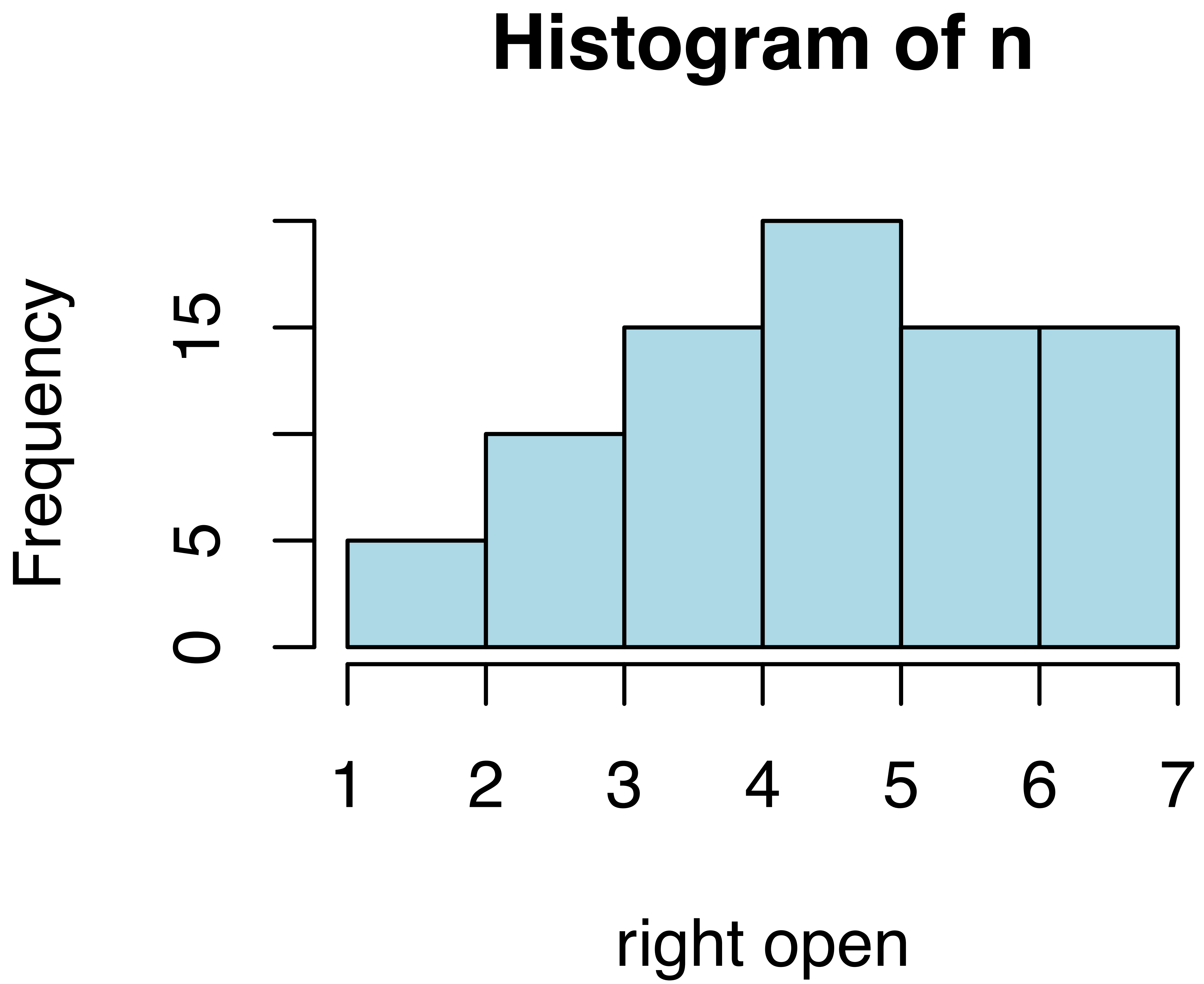
see: ClevelandDotPlot.html

Histograms with discrete data

```
n <- rep(1:7, c(5,10,15,20,15,10,5))
hist(n, col = "lightblue", xlab = "right closed")
```



```r
hist(n, col = "lightblue", right = FALSE, xlab = "right op
```



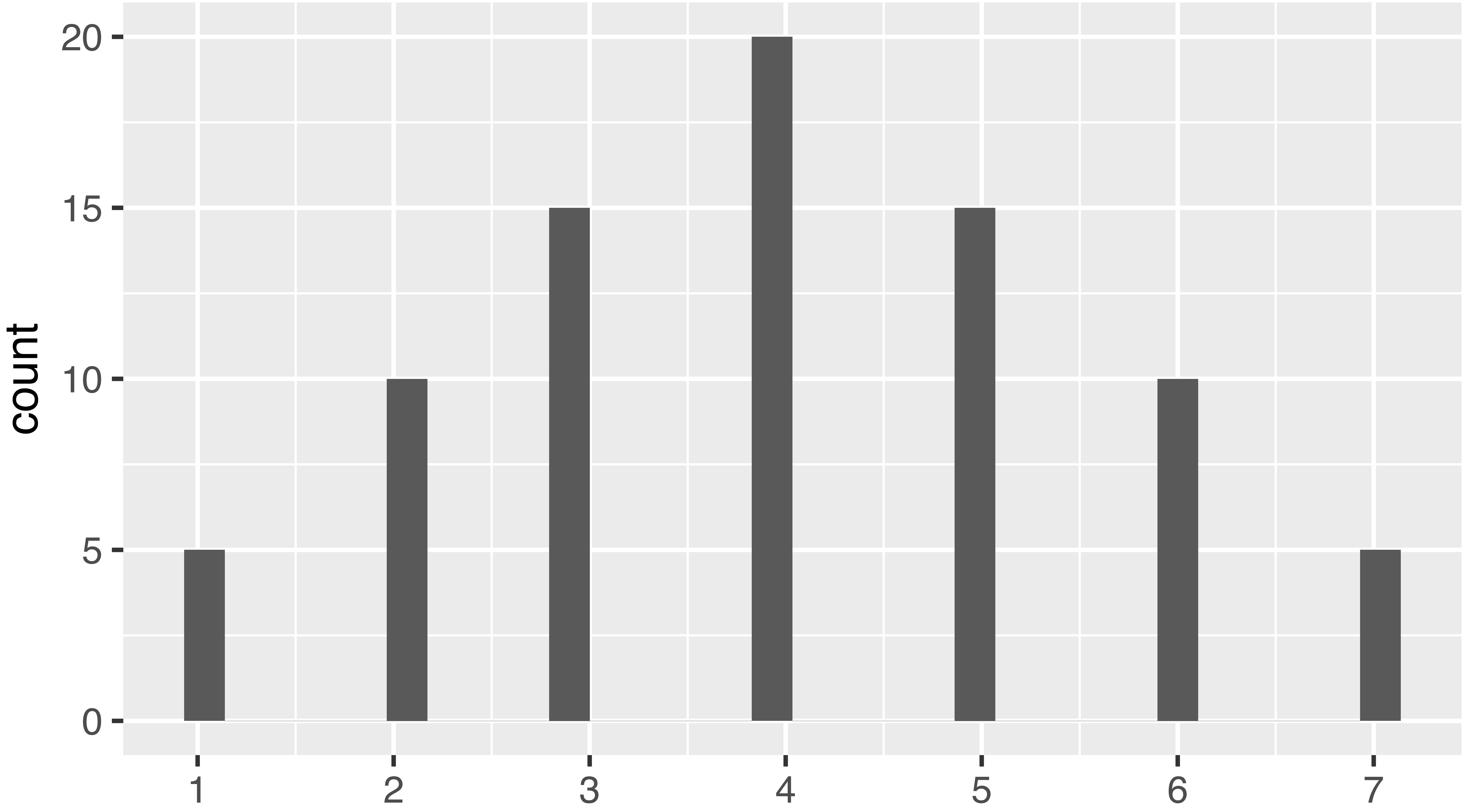
```r
summary(factor(n))
```

```
##  1  2  3  4  5  6  7
##  5 10 15 20 15 10  5
```

Histogram (ggplot2)

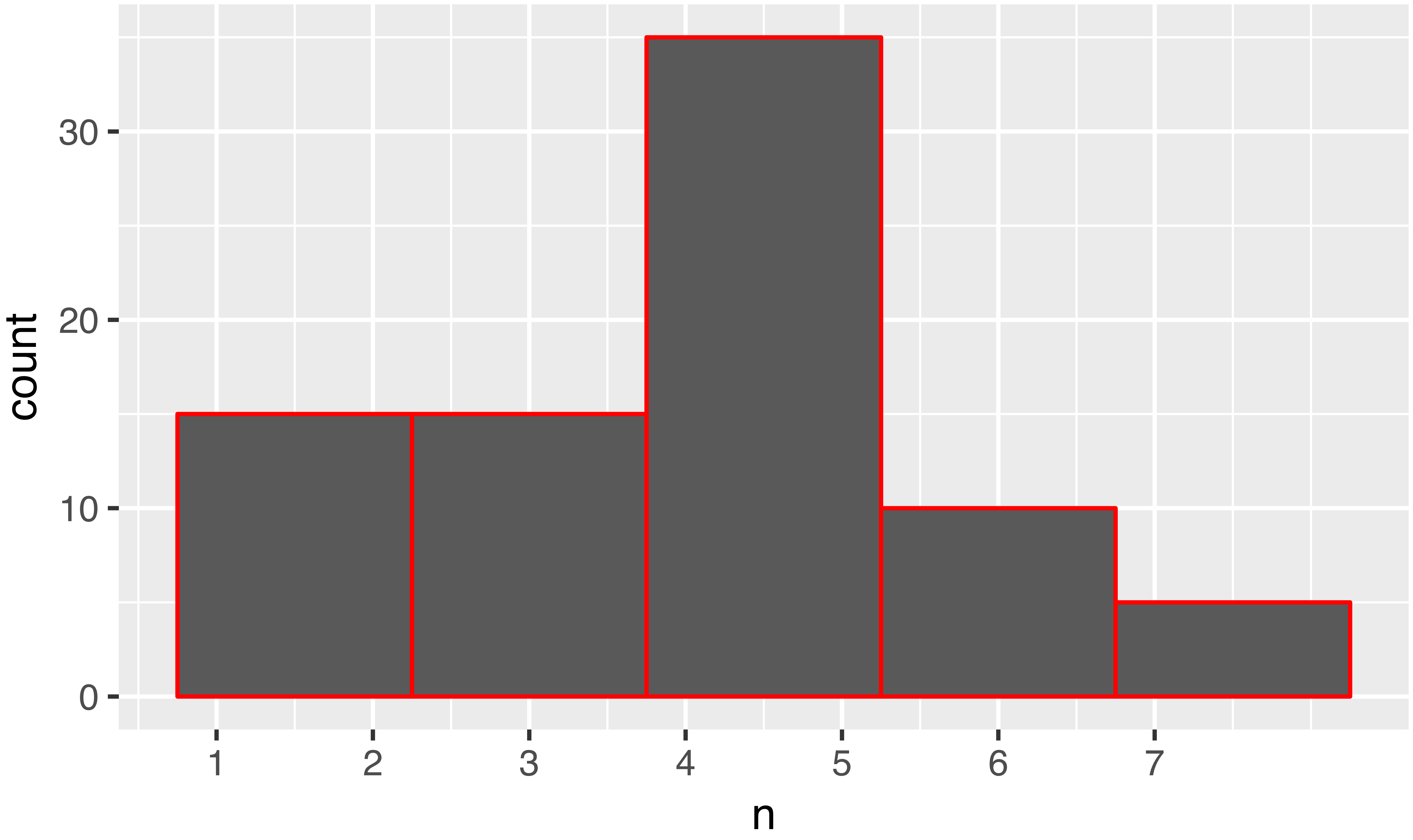
```
df <- data.frame(n)
ggplot(df, aes(x = n)) + geom_histogram() + scale_x_contin
```

```
## `stat_bin()` using `bins = 30`. Pick better value with
```



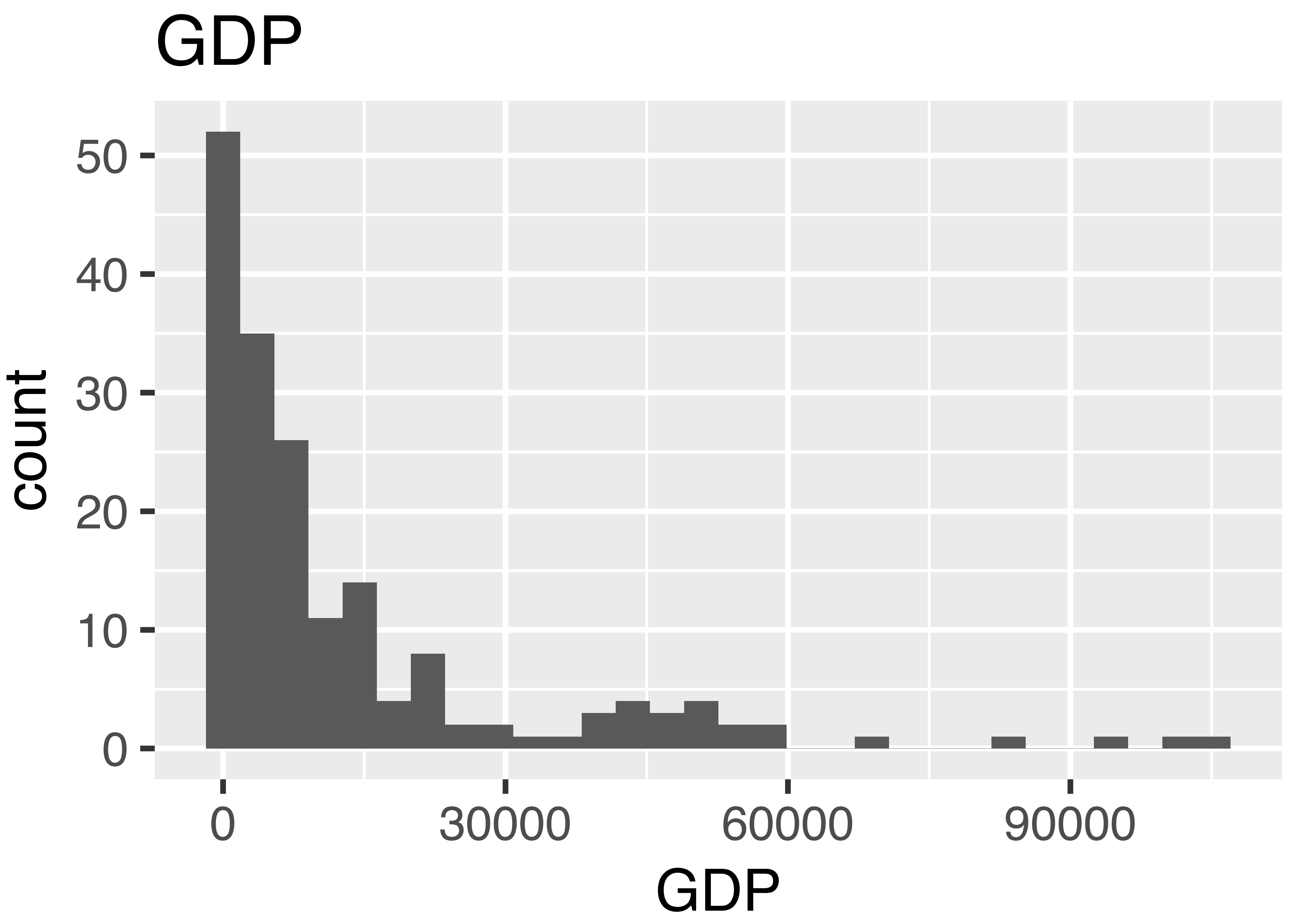
Histograms with discrete data, 5 bins

```
ggplot(df, aes(x = n)) + geom_histogram(bins = 5, color = '
        scale_x_continuous(breaks=1:7)
```



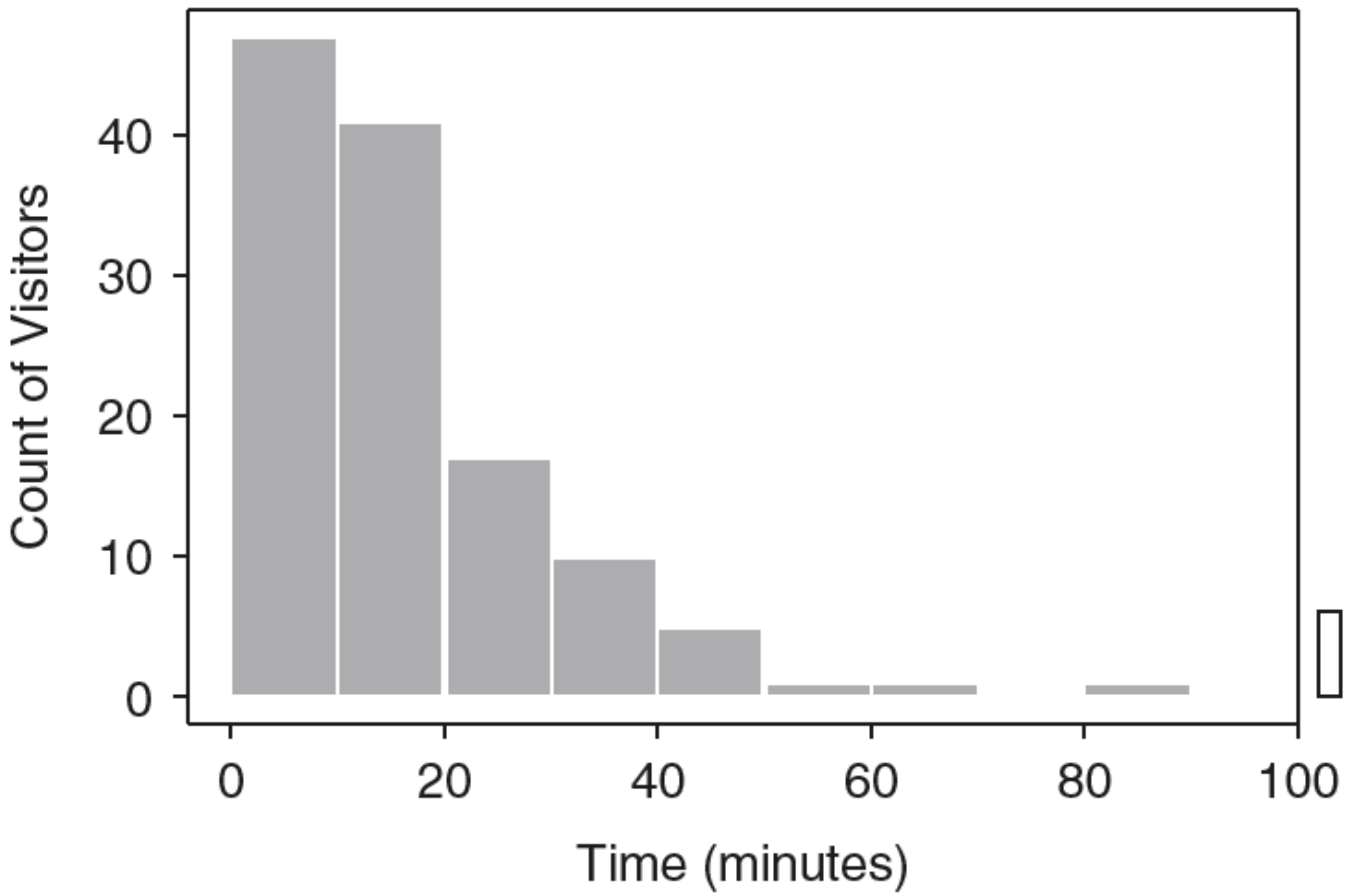
Histogram

```r
ggplot(df, aes(x = GDP)) + geom_histogram() +
    ggtitle ("GDP")
```

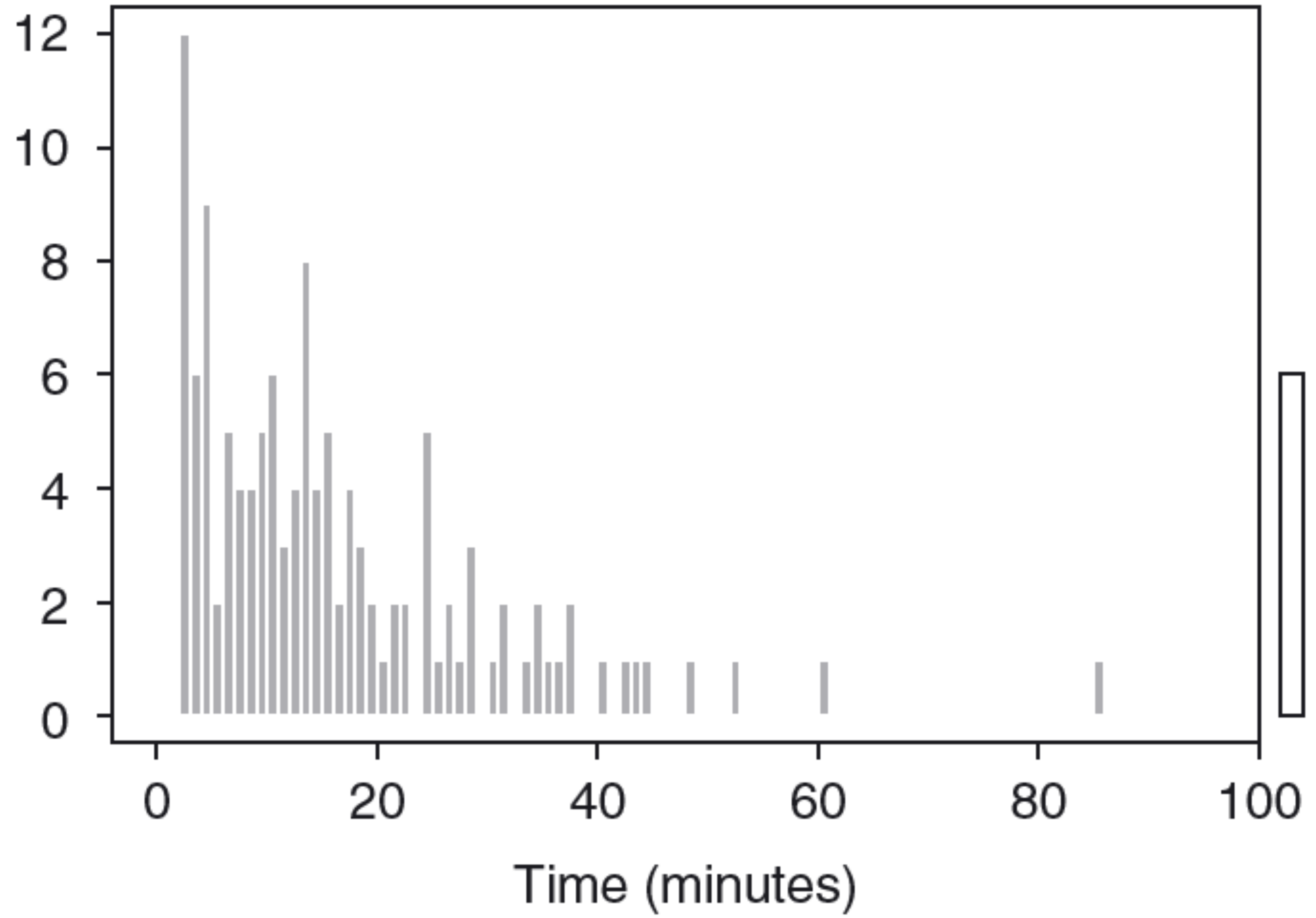


```
`stat_bin()` using `bins = 30`. Pick better value with
  `binwidth`.
```

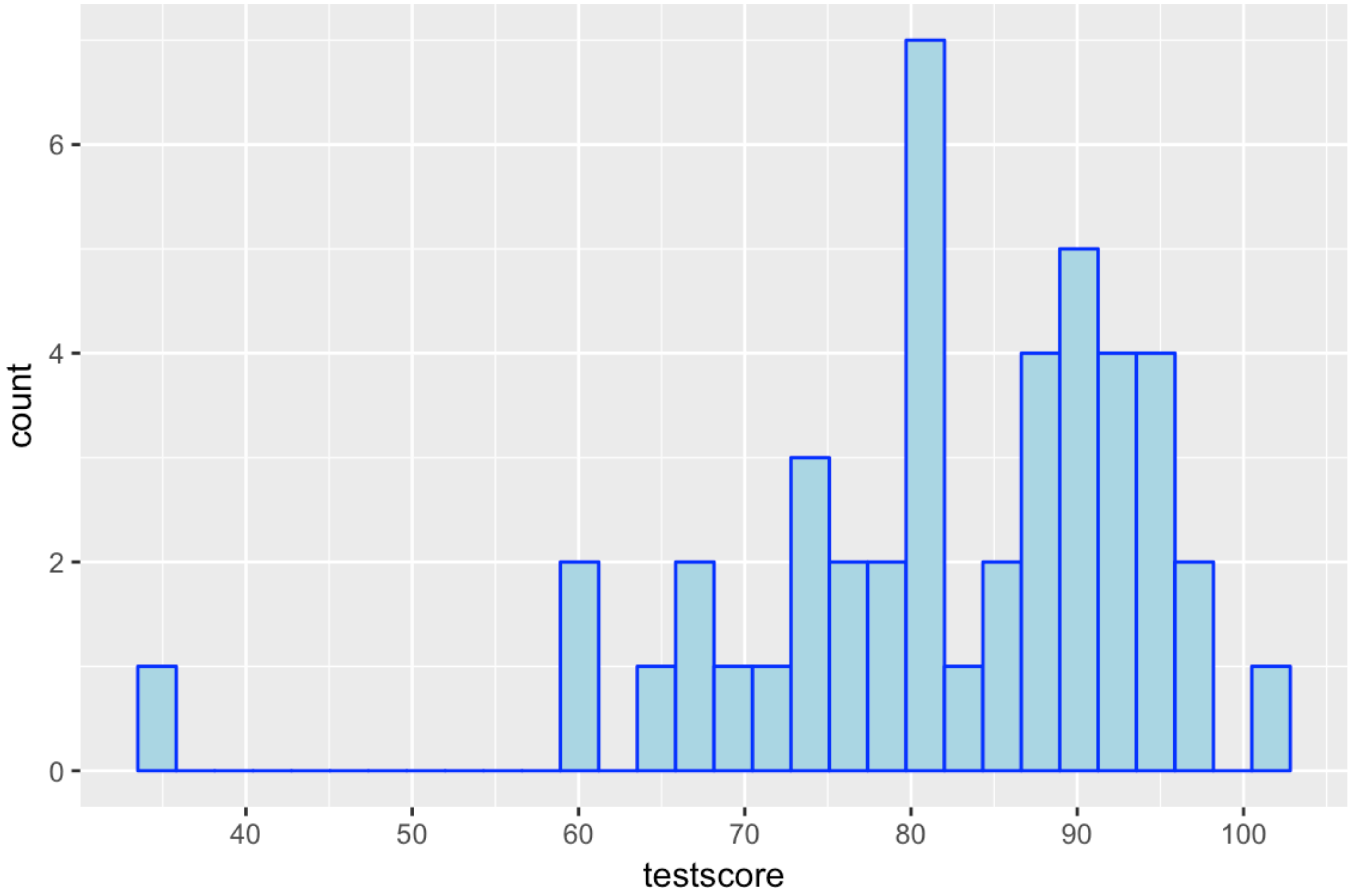
Families Exhibitions Histogram



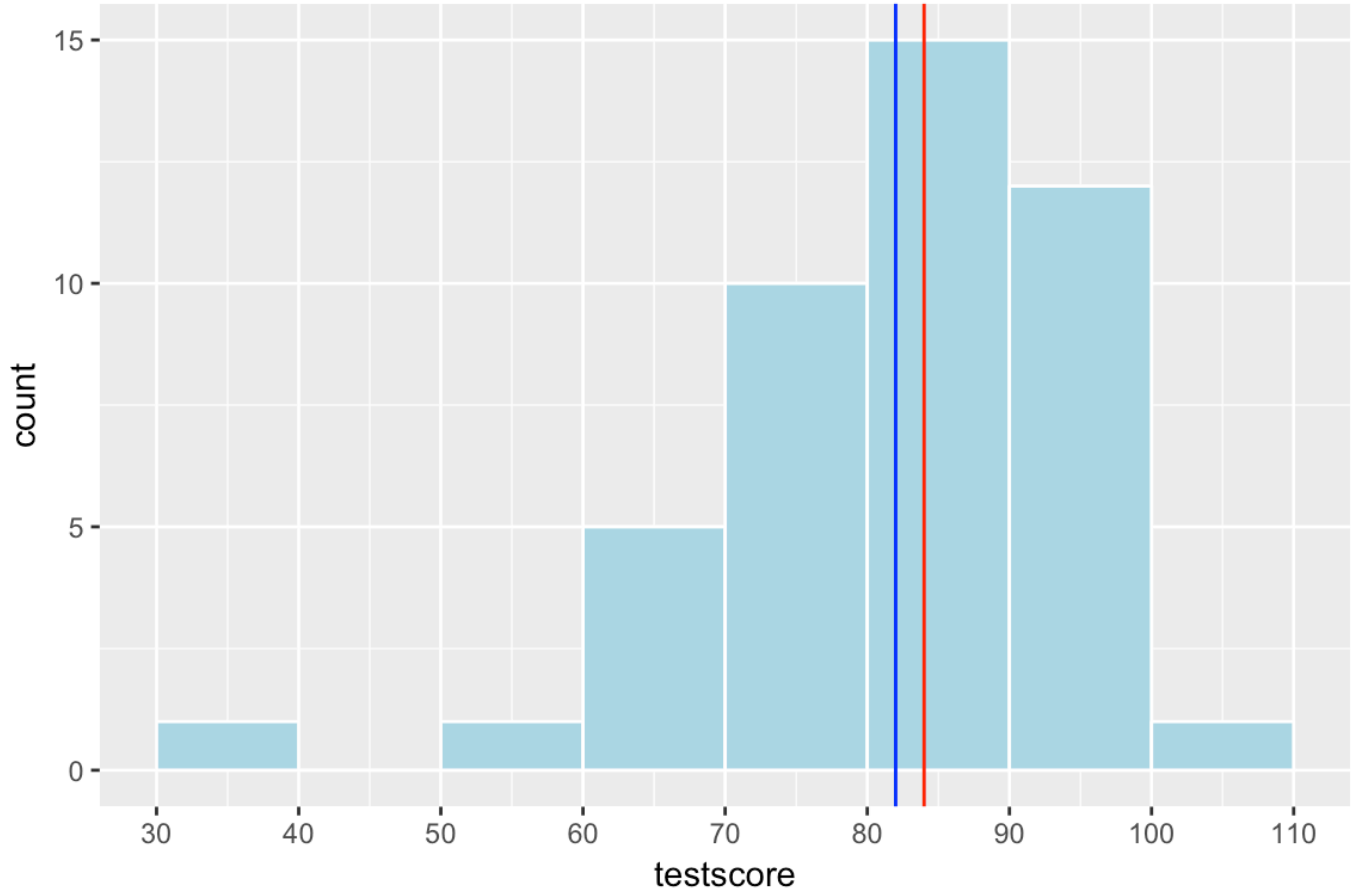
Families Exhibitions Histogram



Test Score Data

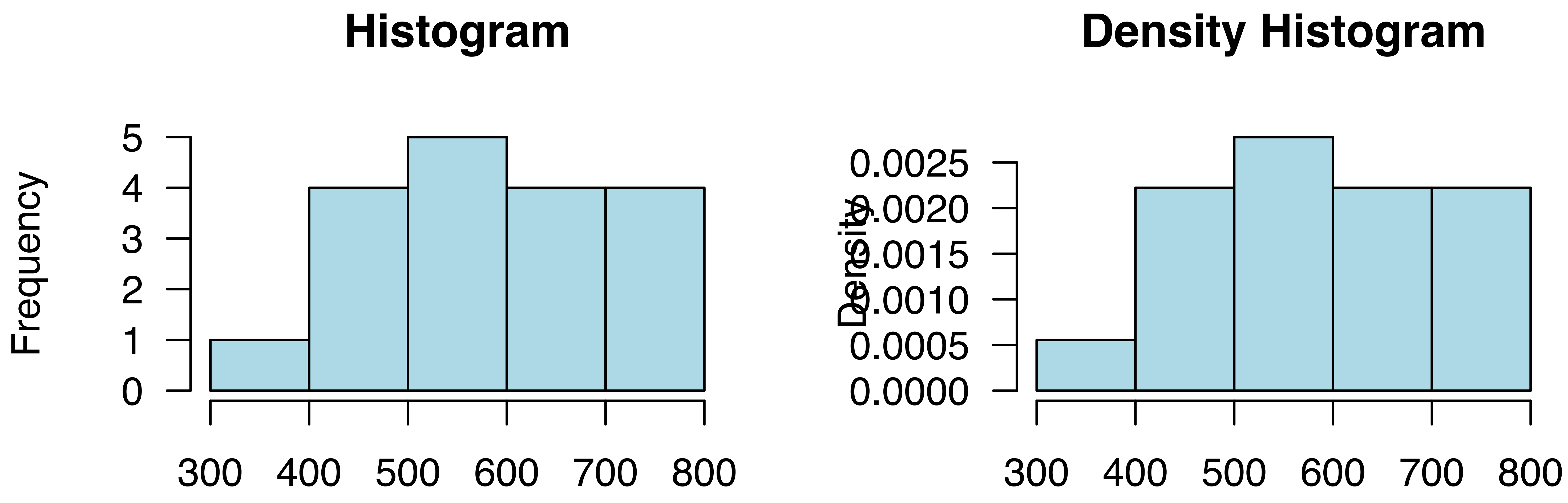


Fewer bins

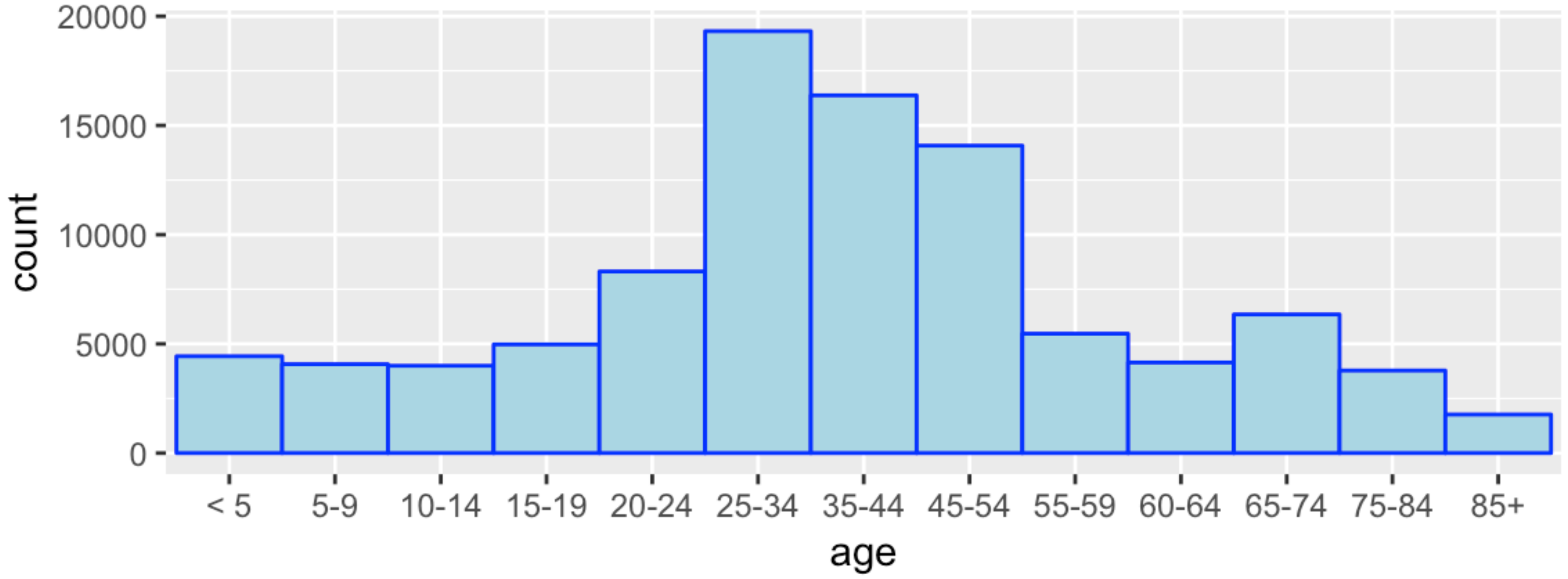


Density histogram

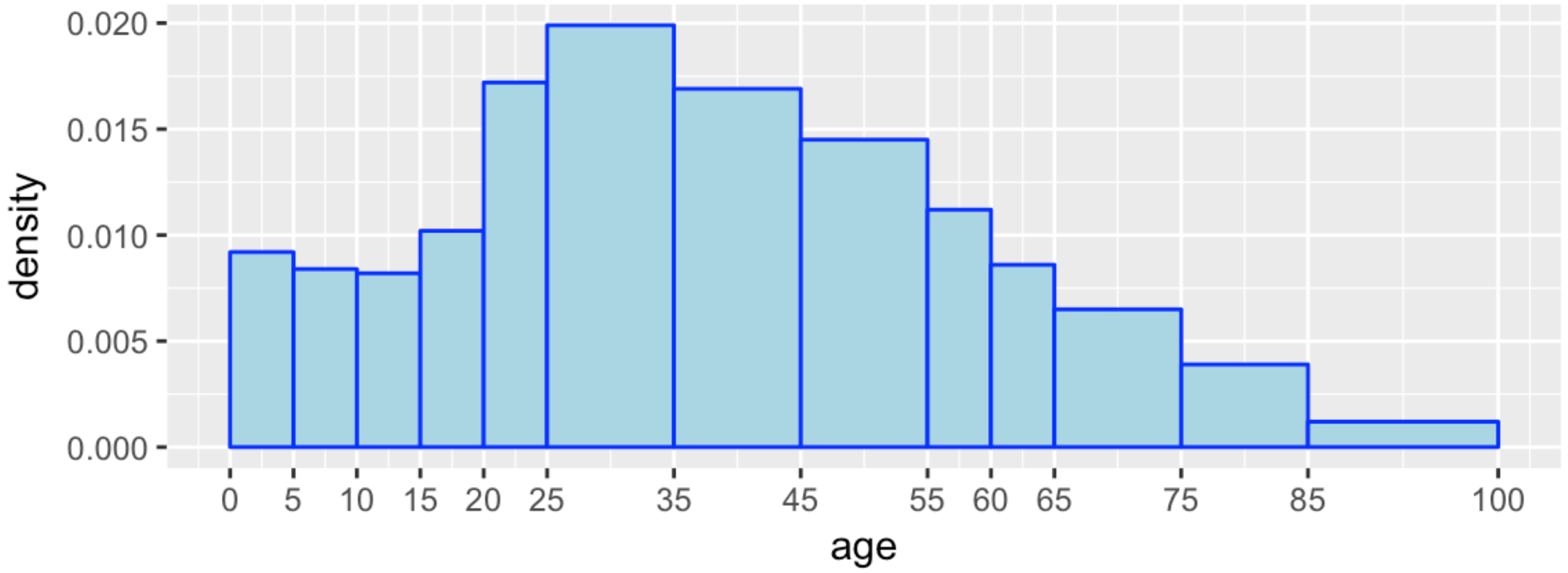
```r
oldpar <- par(mfrow = c(1, 2))
hist(prices, las = 1,
     breaks = c(300, 400, 500, 600, 700, 800),
                col = "lightblue", main = "Histogram")
hist(prices, freq = FALSE, las = 1,
     breaks = c(300, 400, 500, 600, 700, 800),
                col = "lightblue", main =
         "Density Histogram")
```



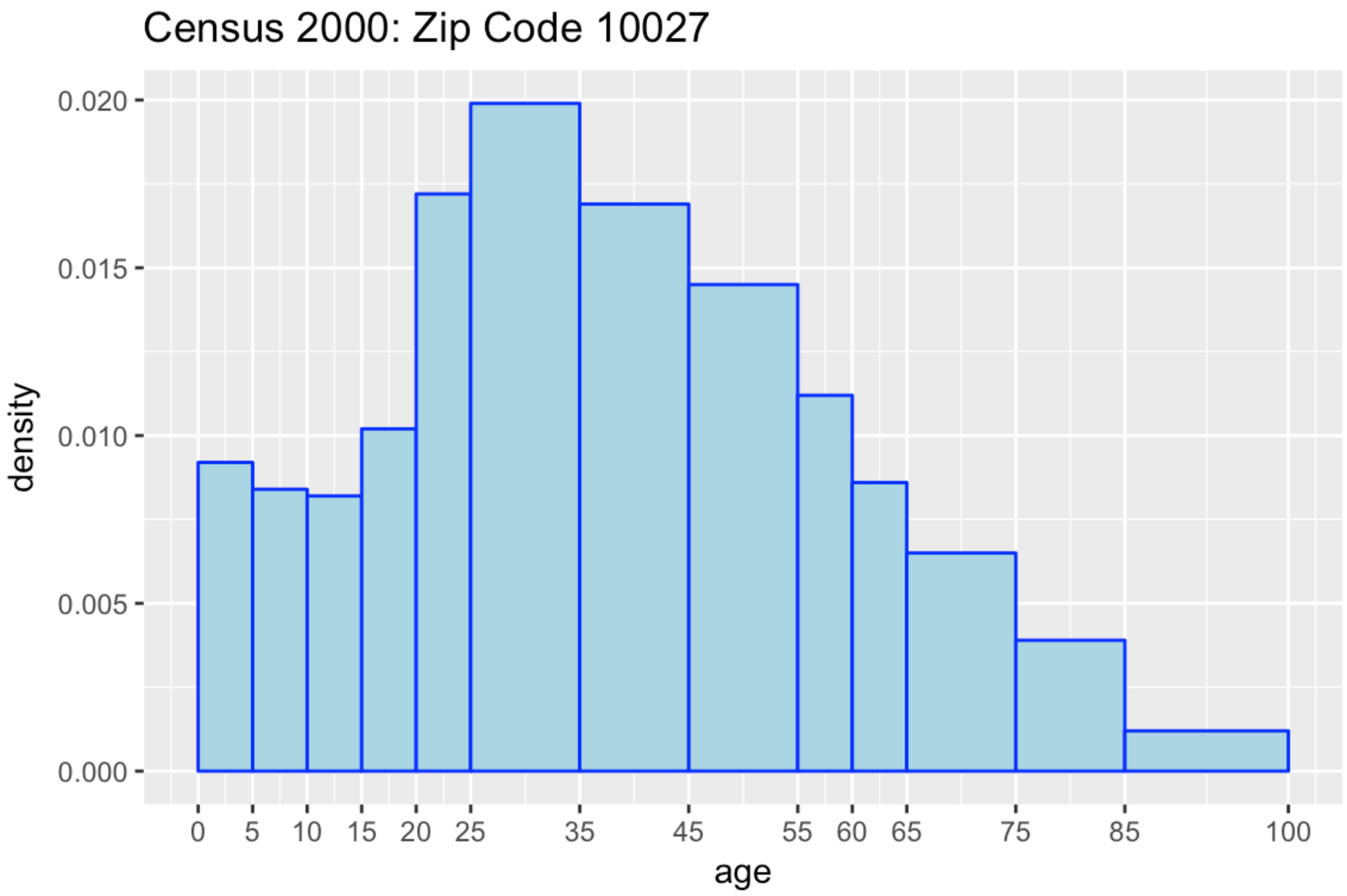
Don't do this.



Do this.



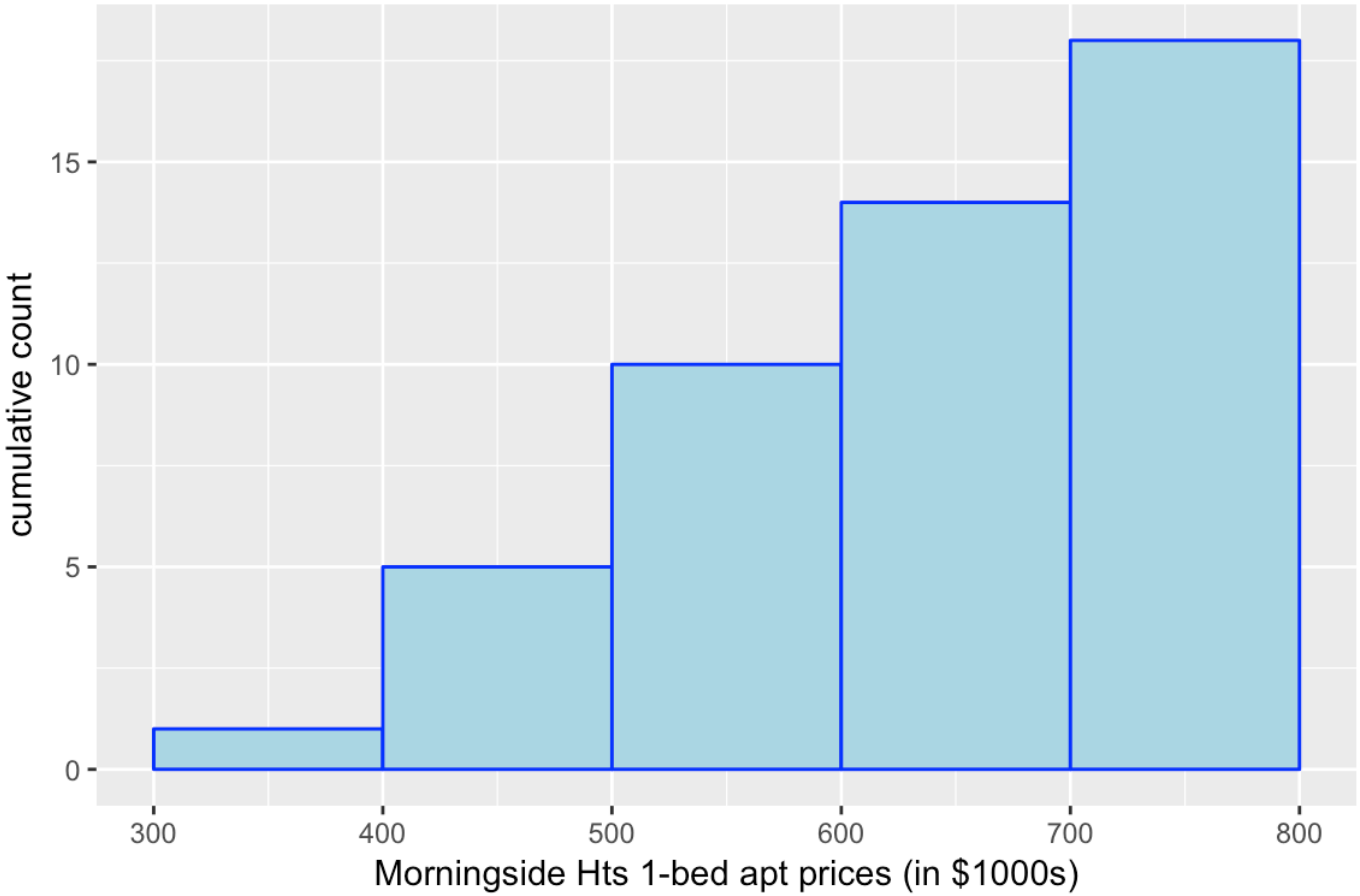
Relative Frequency Histogram with unequal bin (or class) widths



Creating a histogram with unequal class widths

Class	Frequency	RelFreq	ClassWidth	Density
< 5	4435	0.046	5	0.009
5-9	4072	0.042	5	0.008
10-14	3999	0.041	5	0.008
15-19	4977	0.051	5	0.010
20-24	8316	0.086	5	0.017
25-34	19317	0.199	10	0.020
35-44	16380	0.169	10	0.017
45-54	14077	0.145	10	0.014
55-59	5467	0.056	5	0.011
60-64	4148	0.043	5	0.009
65-74	6350	0.065	10	0.007
75-84	3781	0.039	10	0.004
85+	1767	0.018	15	0.001

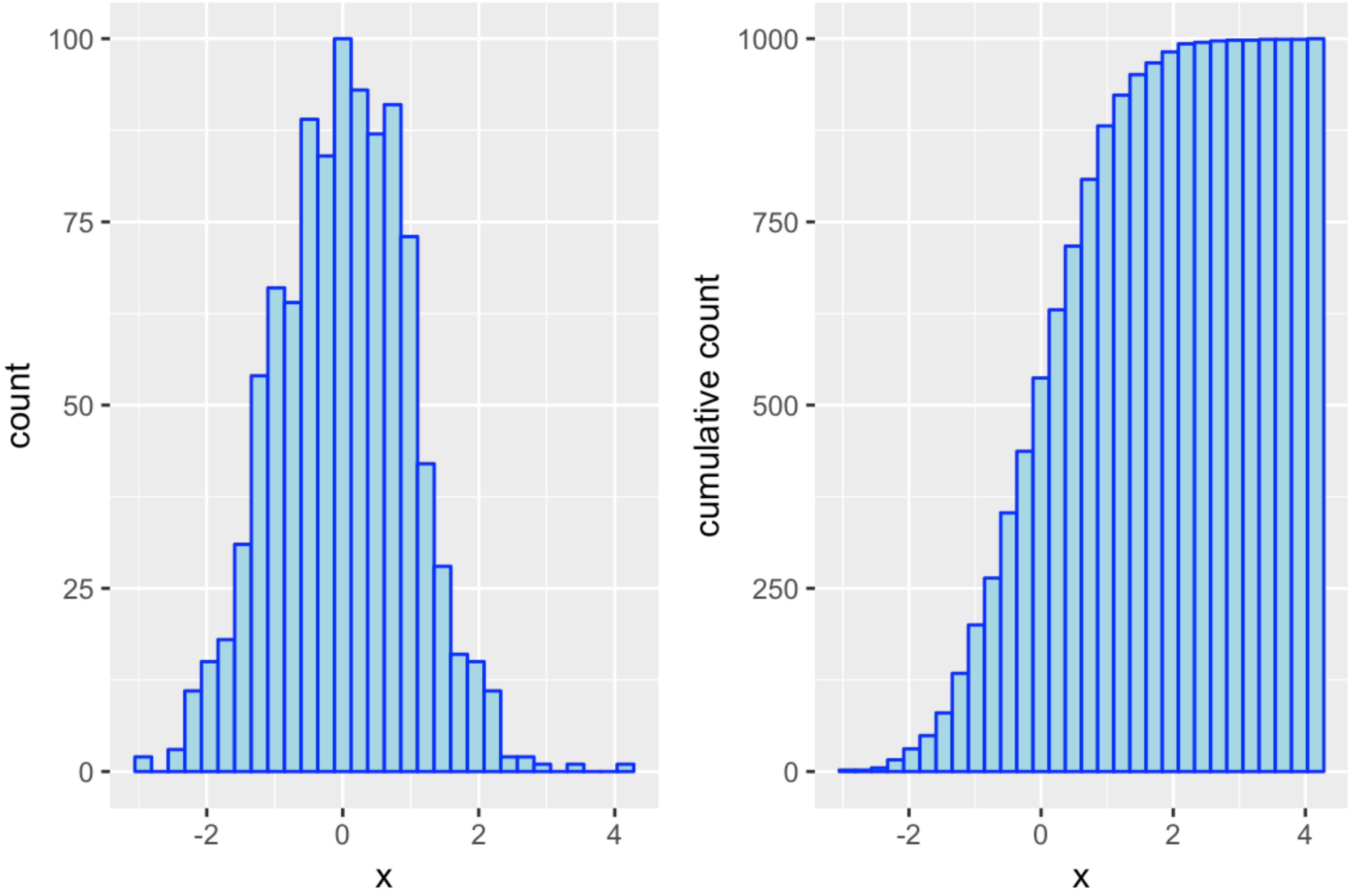
Cumulative Frequency Histogram



Drawing a Cumulative Frequency Histogram

Class	Freq	CumulativeFreq
300-400	1	1
400-500	4	5
500-600	5	10
600-700	4	14
700-800	4	18

Cumulative Frequency Histogram



Morningside Heights 1-bedroom apt. prices (in 1000s)

```r
prices <- c(379, 425, 450, 450, 499, 529, 535, 535, 545,
            599, 665, 675, 699, 699, 725, 725, 745, 799)
```

Stem and leaf plot

```r
signif(prices, 2)
```

```
##  [1] 380 420 450 450 500 530 540 540 540 600 660 680 70
## [18] 800
```

```r
stem(prices)
```

```
## 
##   The decimal point is 2 digit(s) to the right of the |
## 
##   3 | 8
##   4 | 355
##   5 | 03445
##   6 | 078
##   7 | 00335
##   8 | 0
```

Five number summary

1. minimum
2. 1st quartile
3. middle number (median)
4. 3rd quantile
5. maximum

```r
quantile(prices)
```

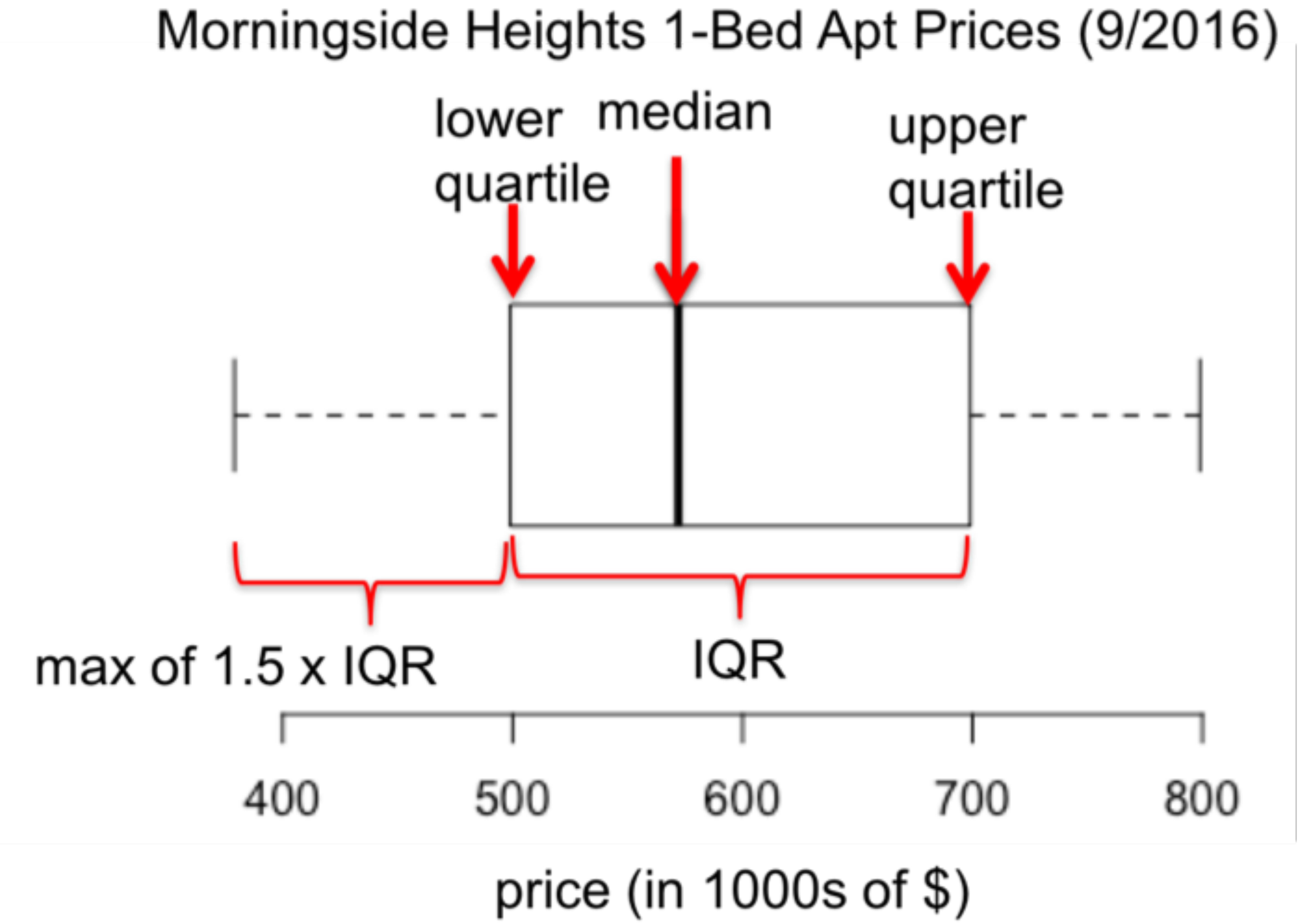
```
##    0%   25%   50%   75%  100%
## 379.0 506.5 572.0 699.0 799.0
```

Quantile Methods

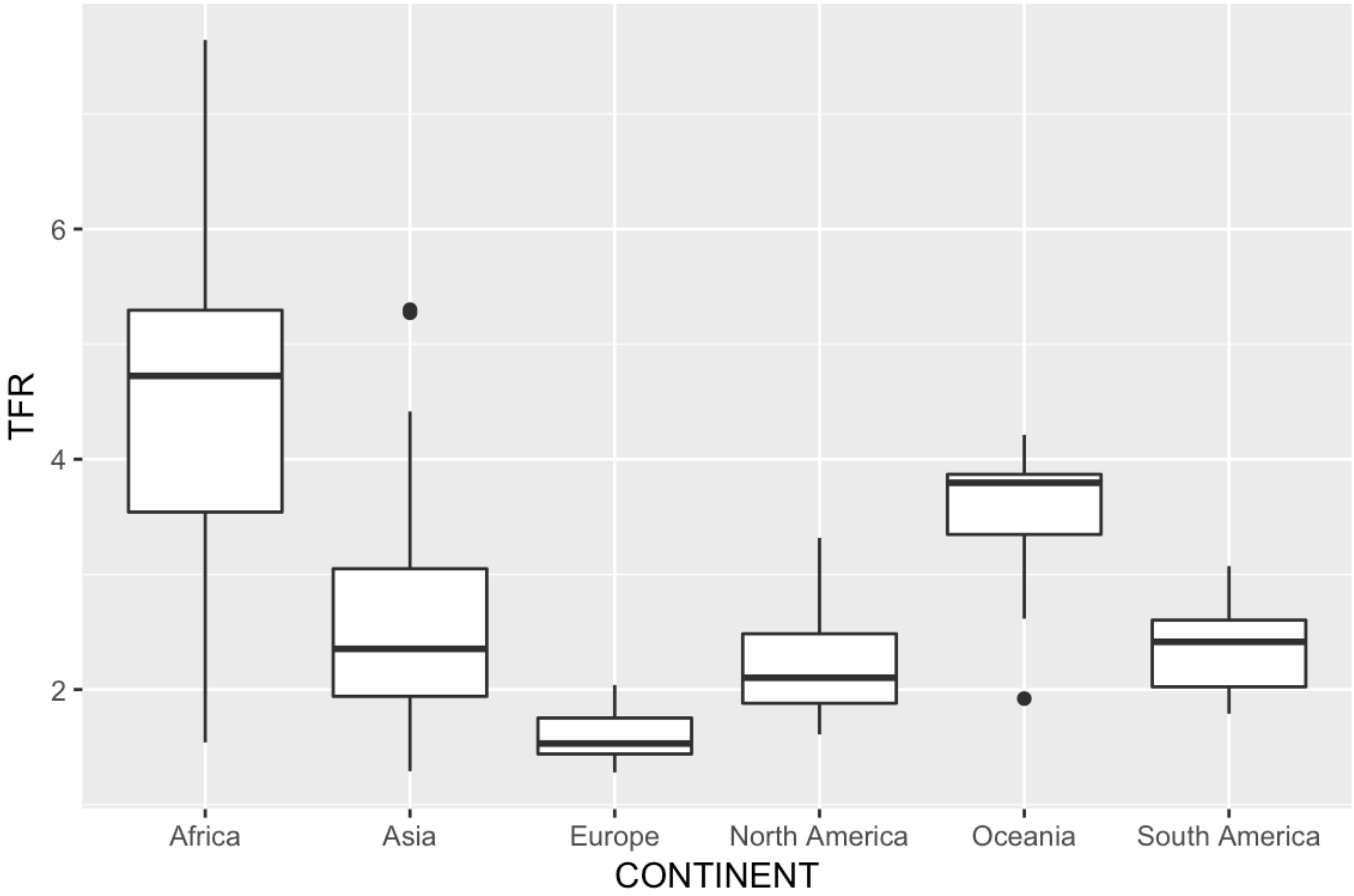
```r
q <- matrix(,nrow = 9, ncol = 5)
for (i in 1:9) q[i,] <- quantile(gdp, type = i)
q <- data.frame(q)
colnames(q) = c("min", "Q1", "med", "Q3", "max")
```

	min	Q1	med	Q3	max
	244.1965	1586.780	5583.616	14357.41	105447.1
	244.1965	1586.780	5583.616	14357.41	105447.1
	244.1965	1586.780	5583.616	14342.52	105447.1
	244.1965	1562.097	5557.696	14346.25	105447.1
	244.1965	1600.384	5583.616	14353.69	105447.1
	244.1965	1586.780	5583.616	14357.41	105447.1
→	244.1965	1613.989	5583.616	14349.97	105447.1
	244.1965	1595.850	5583.616	14354.93	105447.1
	244.1965	1596.983	5583.616	14354.62	105447.1

Box plot


Morningside Heights 1-Bed Apt Prices (9/2016)
lower quartile
median
upper quartile
max of 1.5 x IQR
IQR
400
500
600
700
800
price (in 1000s of $)

Multiple box plots


TFR
Africa
Asia
Europe
North America
Oceania
South America
CONTINENT