# Heatmaps

# Heatmaps

- Can show frequency counts (2D histogram) or value of a third variable
- Can be used for continuous or categorical data

# History

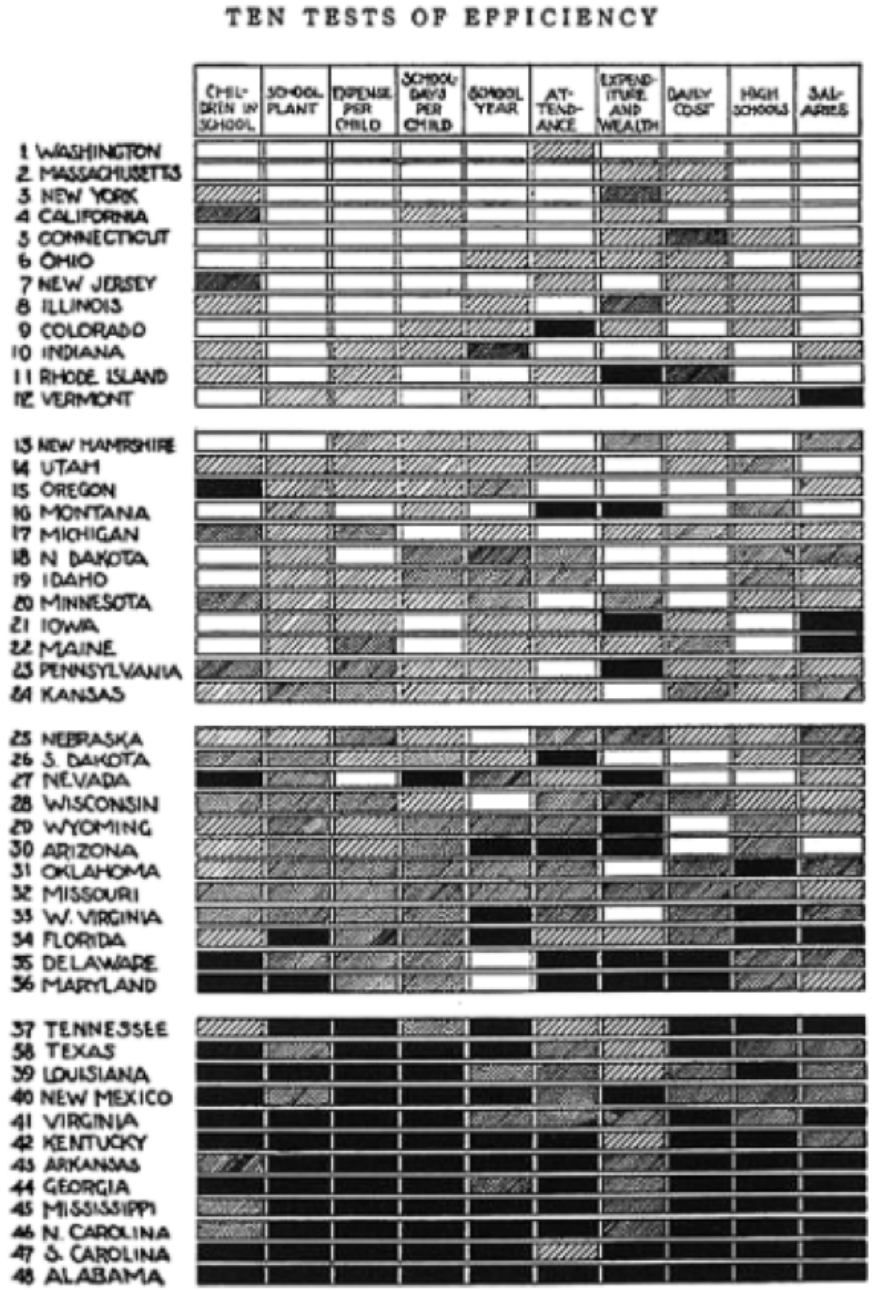

TEN TESTS OF EFFICIENCY
CHILDREN IN SCHOOL
SCHOOL PLANT
EXPENSE PER CHILD
SCHOOL DAYS PER CHILD
SCHOOL YEAR
ATTENDANCE
EXPENDITURE AND WEALTH
DAILY COST
HIGH SCHOOLS
SALARIES
1 WASHINGTON
2 MASSACHUSETTS
3 NEW YORK
4 CALIFORNIA
5 CONNECTICUT
6 OHIO
7 NEW JERSEY
8 ILLINOIS
9 COLORADO
10 INDIANA
11 RHODE ISLAND
12 VERMONT
13 NEW HAMPSHIRE
14 UTAH
15 OREGON
16 MONTANA
17 MICHIGAN
18 N. DAKOTA
19 IDAHO
20 MINNESOTA
21 IOWA
22 MAINE
23 PENNSYLVANIA
24 KANSAS
25 NEBRASKA
26 S. DAKOTA
27 NEVADA
28 WISCONSIN
29 WYOMING
30 ARIZONA
31 OKLAHOMA
32 MISSOURI
33 W. VIRGINIA
34 FLORIDA
35 DELAWARE
36 MARYLAND
37 TENNESSEE
38 TEXAS
39 LOUISIANA
40 NEW MEXICO
41 VIRGINIA
42 KENTUCKY
43 ARKANSAS
44 GEORGIA
45 MISSISSIPPI
46 N. CAROLINA
47 S. CAROLINA
48 ALABAMA

Figure 3. Sorted shaded display from Brinton (1914). The data are ranks of U.S. states on each of 10 educational features assessed in 1910. The matrix has been sorted by the row-marginal ranks.

# History

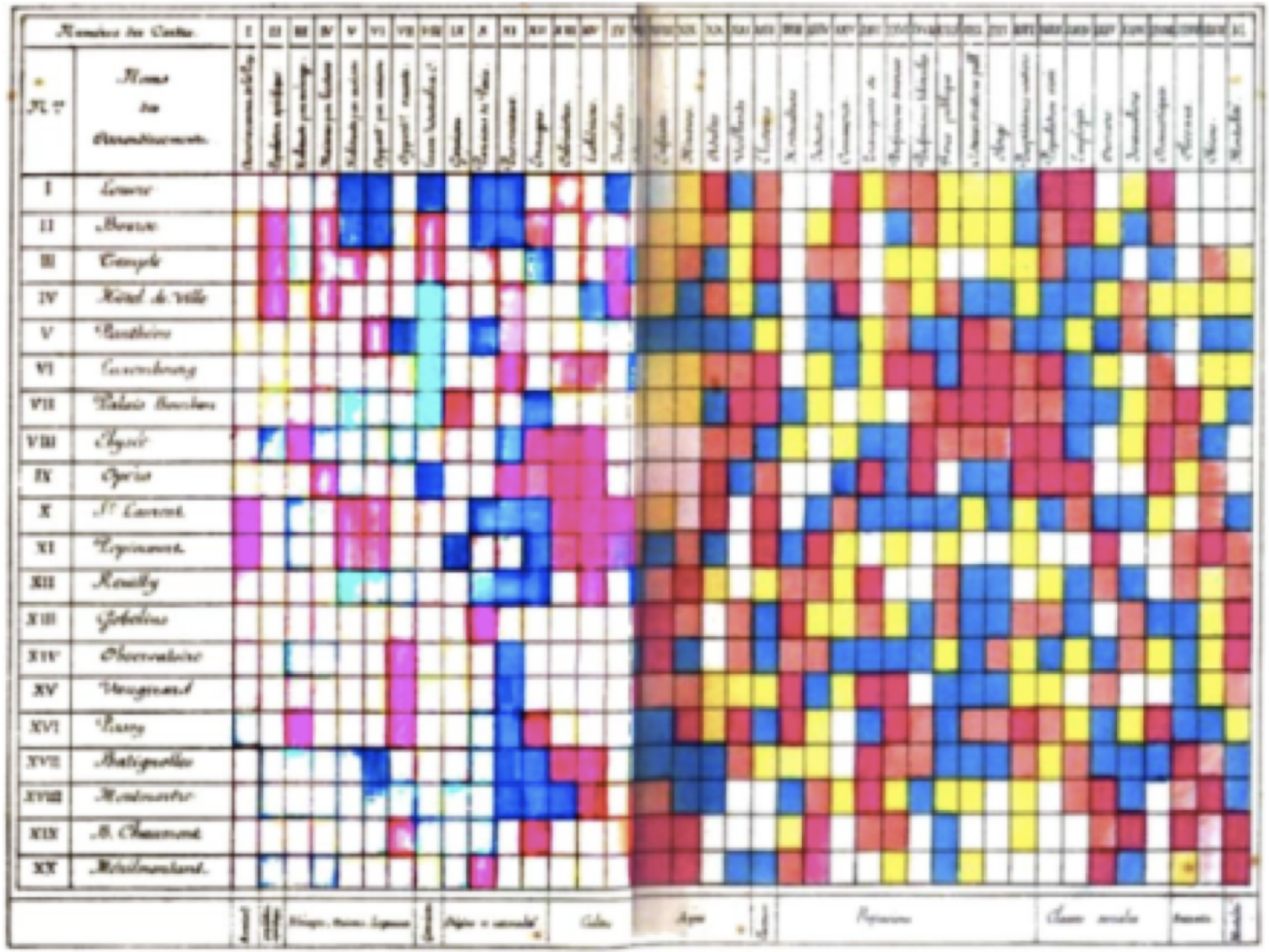
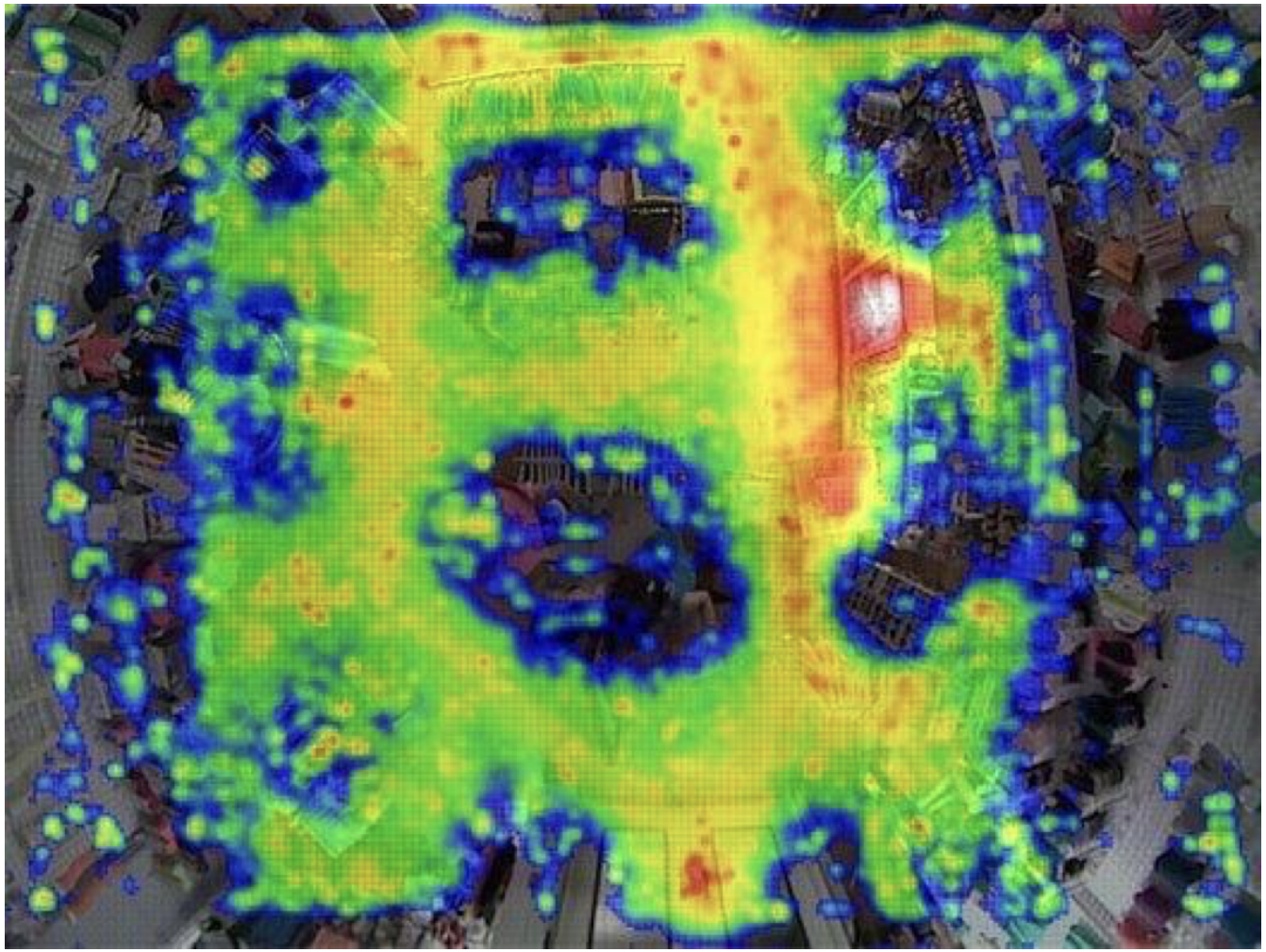


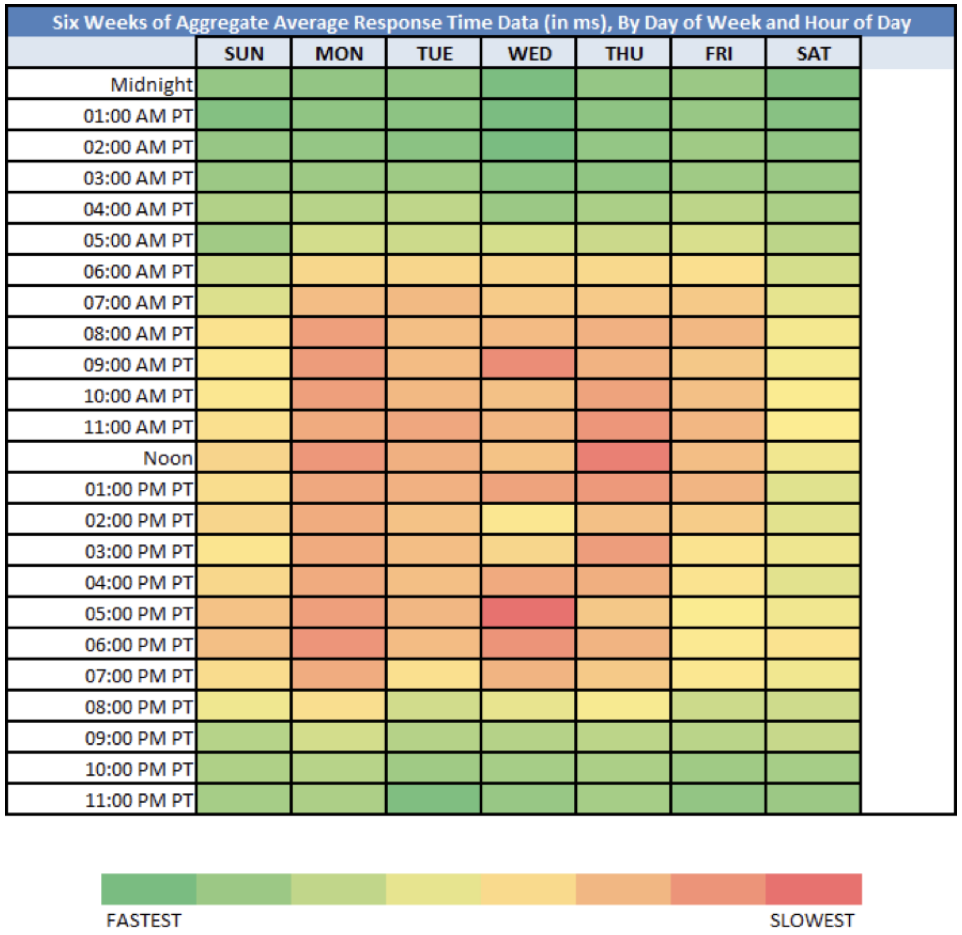
Figure 2. Shaded matrix display from Loua (1873), available online at http://books.google.com/books/. This was designed as a summary of 40 separate maps of Paris, showing the characteristics (e.g., national origin, professions, age, social classes) of 20 districts, using a color scale ranging from white (low) through yellow and blue to red (high).

# In-store shopping pattern



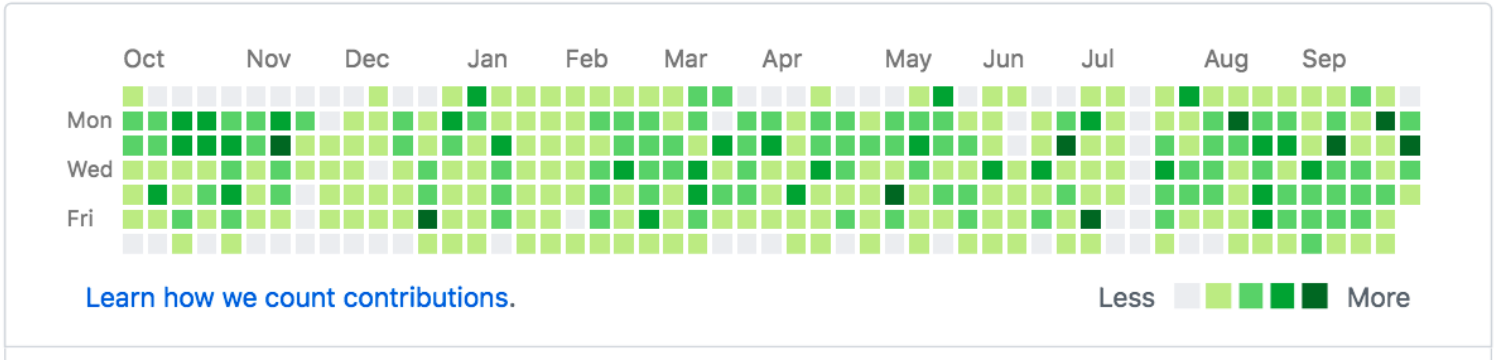
# Eye-tracking studies

# Demand

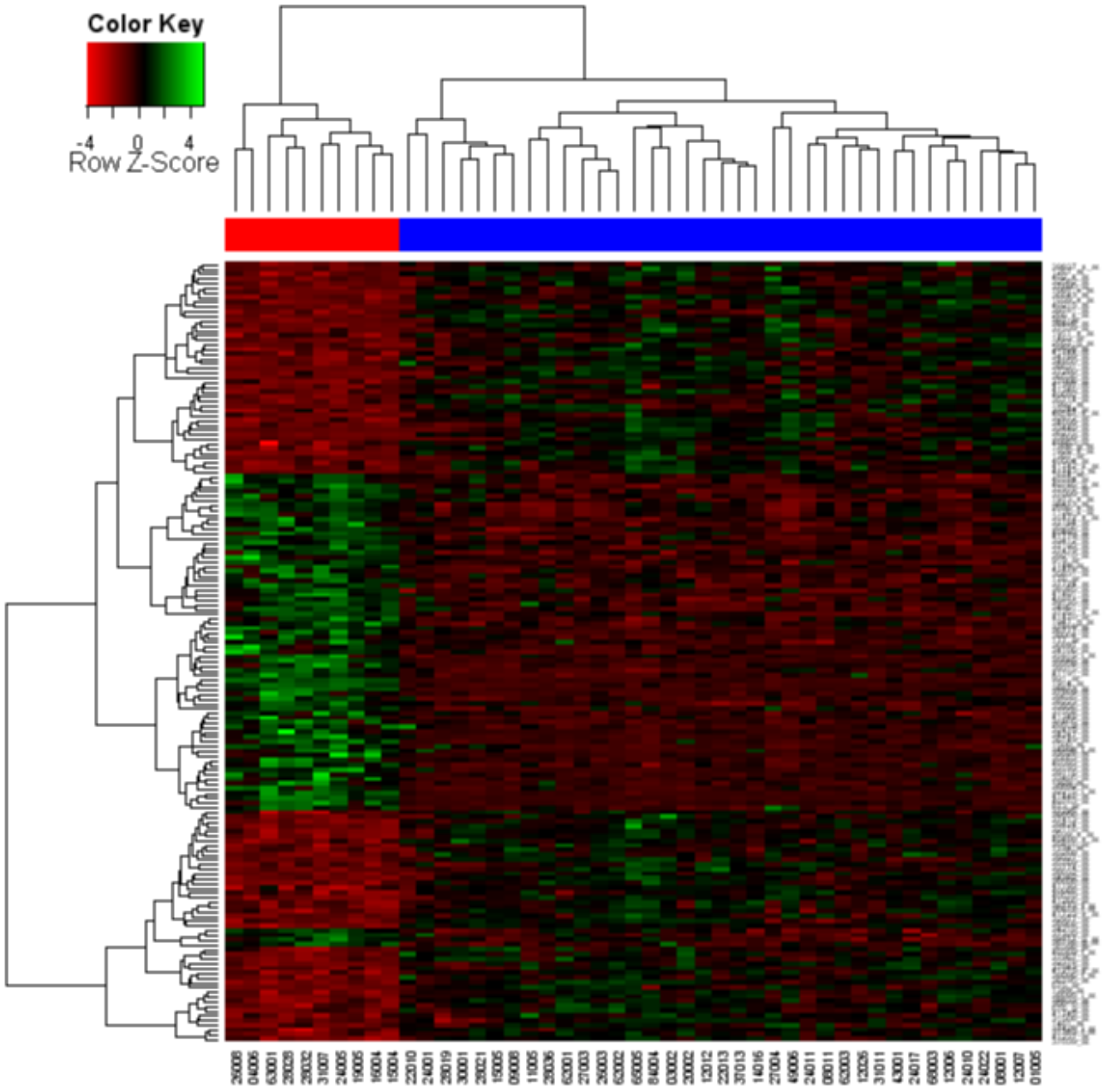

Six Weeks of Aggregate Average Response Time Data (in ms), By Day of Week and Hour of Day
SUN
MON
TUE
WED
THU
FRI
SAT
Midnight
01:00 AM PT
02:00 AM PT
03:00 AM PT
04:00 AM PT
05:00 AM PT
06:00 AM PT
07:00 AM PT
08:00 AM PT
09:00 AM PT
10:00 AM PT
11:00 AM PT
Noon
01:00 PM PT
02:00 PM PT
03:00 PM PT
04:00 PM PT
05:00 PM PT
06:00 PM PT
07:00 PM PT
08:00 PM PT
09:00 PM PT
10:00 PM PT
11:00 PM PT
FASTEST
SLOWEST

# Frequency counts

4,133 contributions in the last year



# Gene Expression



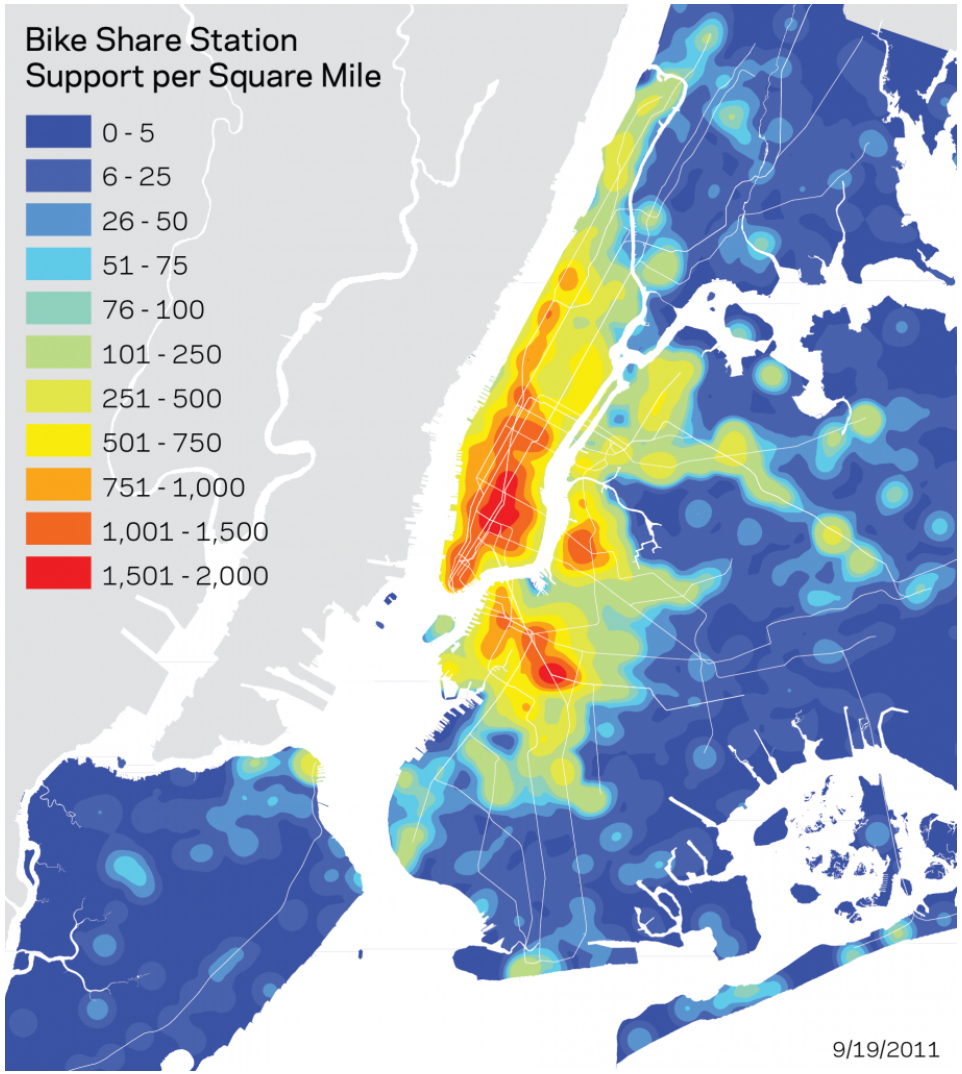
rows: genes

columns: samples

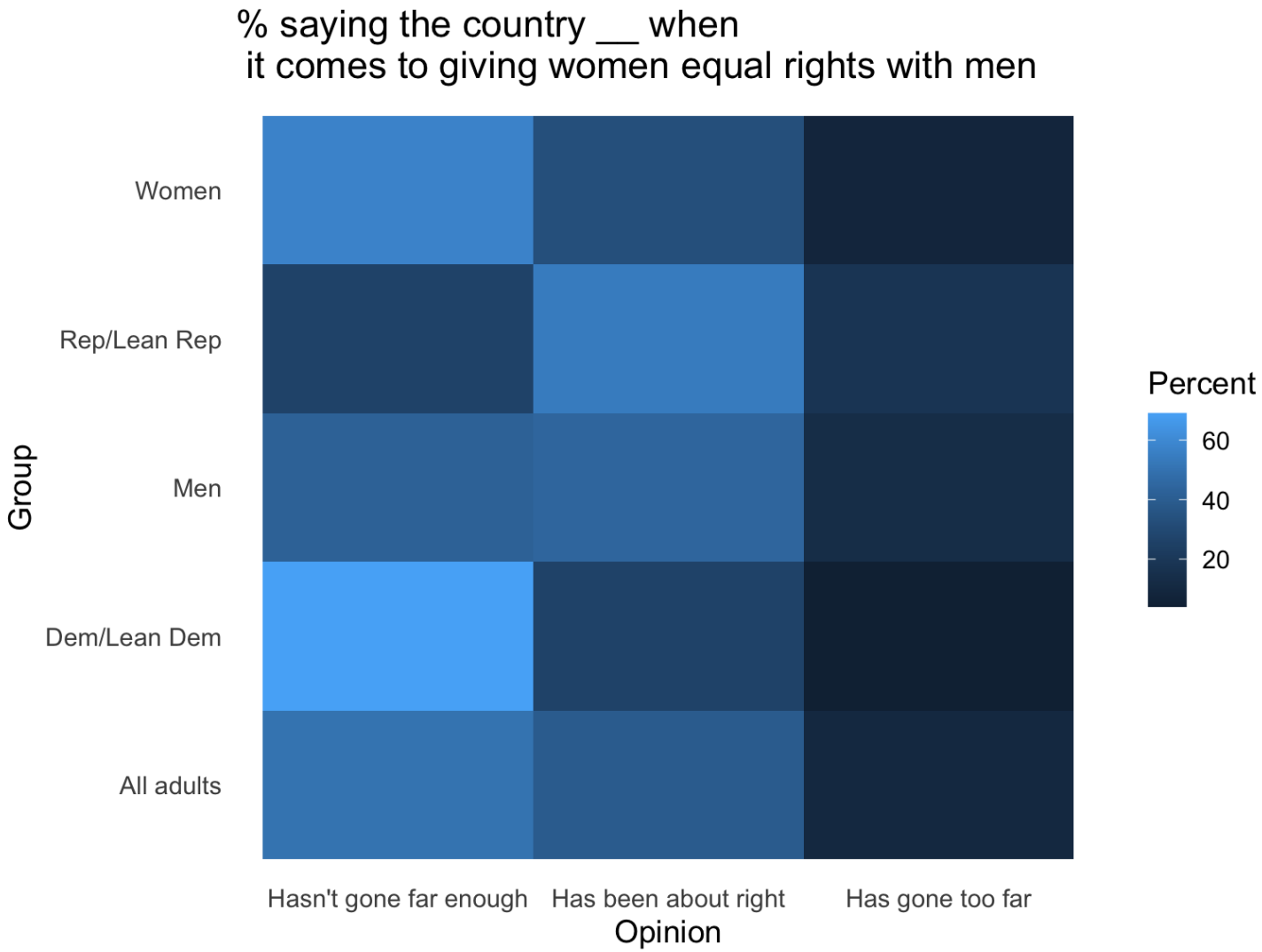
color: change in gene expression level

Source: https://warwick.ac.uk/fac/sci/moac/people/students/peter_cock/r/heatmap/

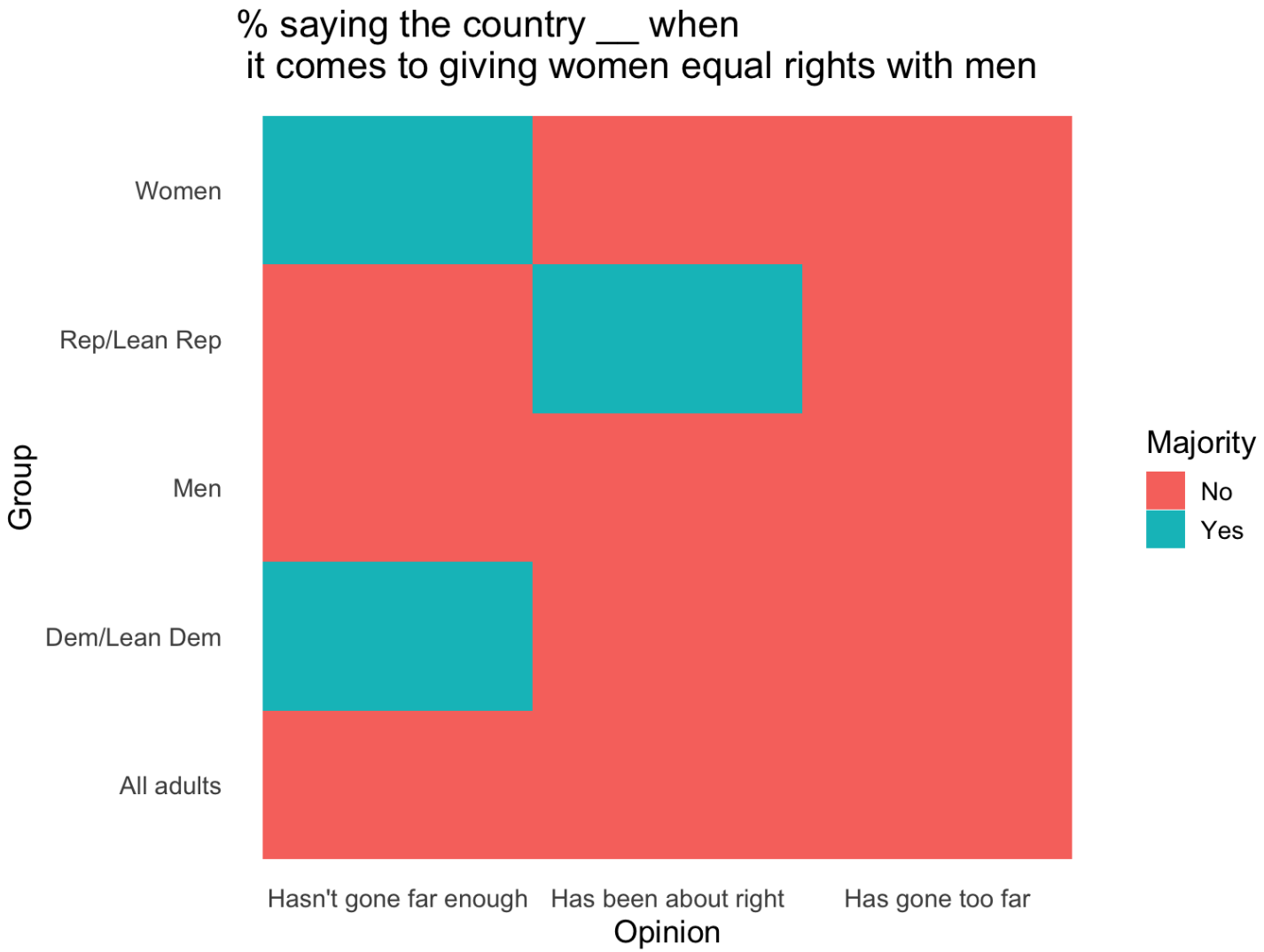
# Geographic pattern



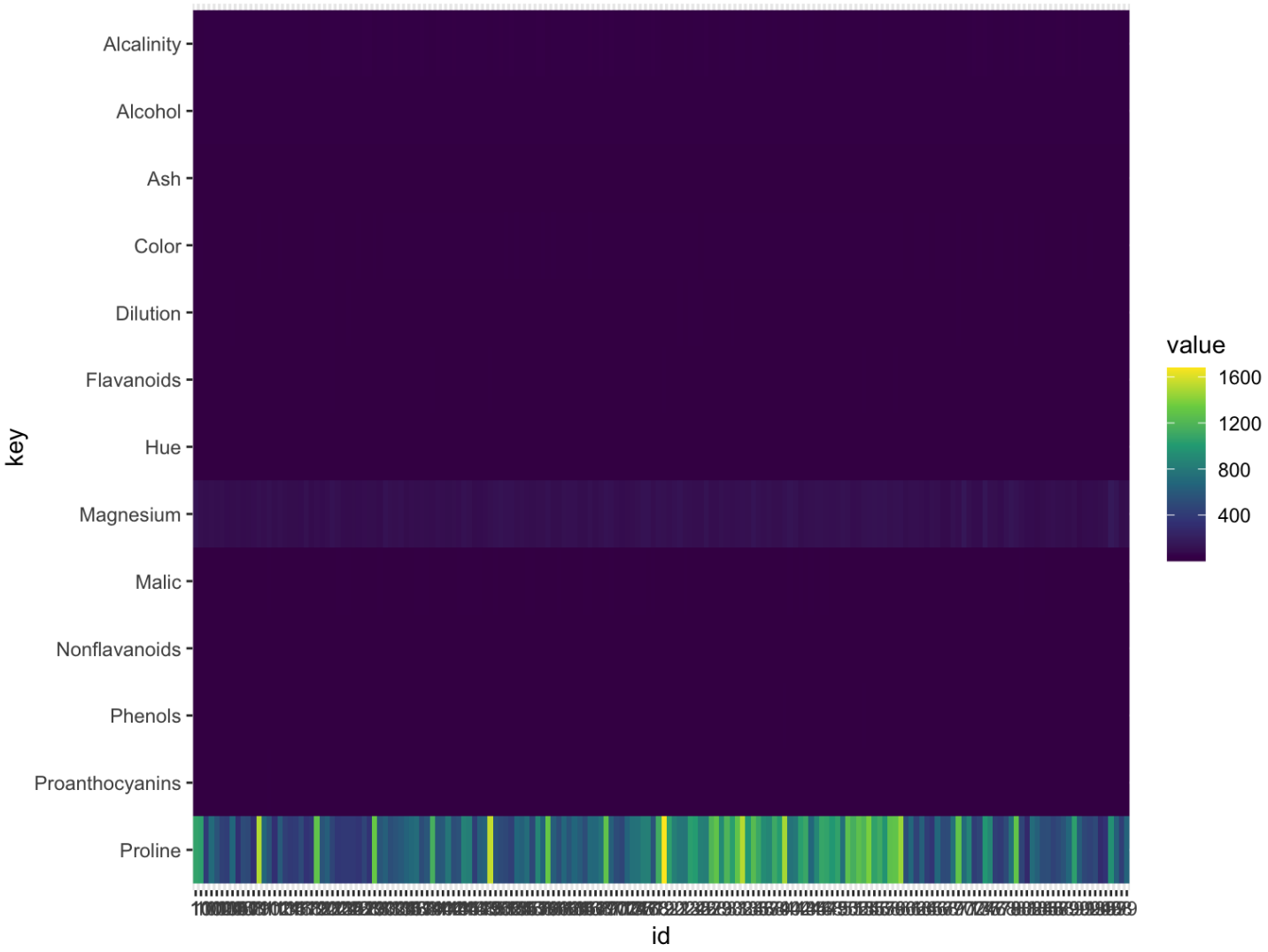
# heat maps

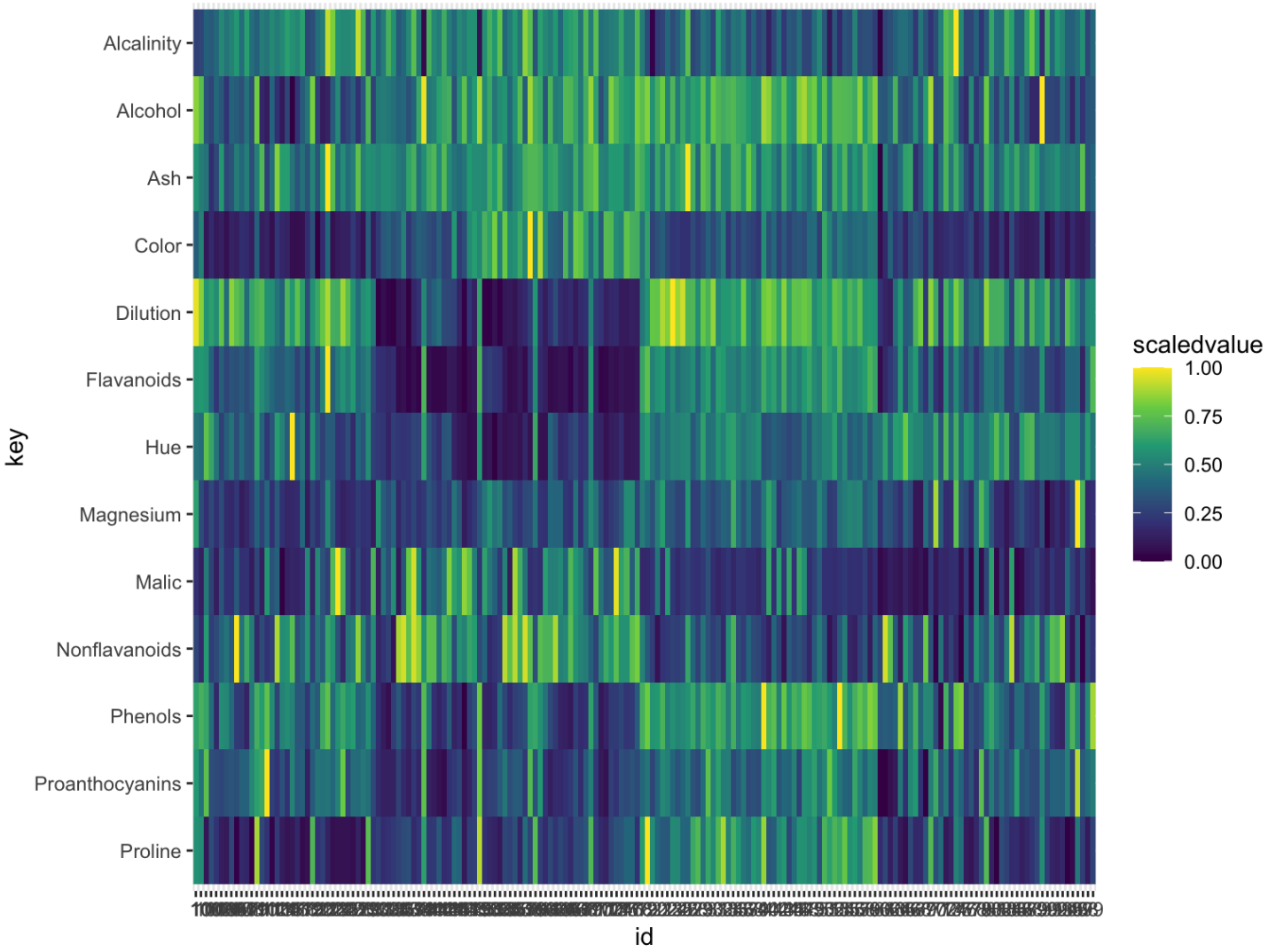

% saying the country __ when
it comes to giving women equal rights with men
Women
Rep/Lean Rep
Men
Dem/Lean Dem
All adults
Group
Hasn't gone far enough
Has been about right
Has gone too far
Opinion
Percent
60
40
20

# numerical –> categorical

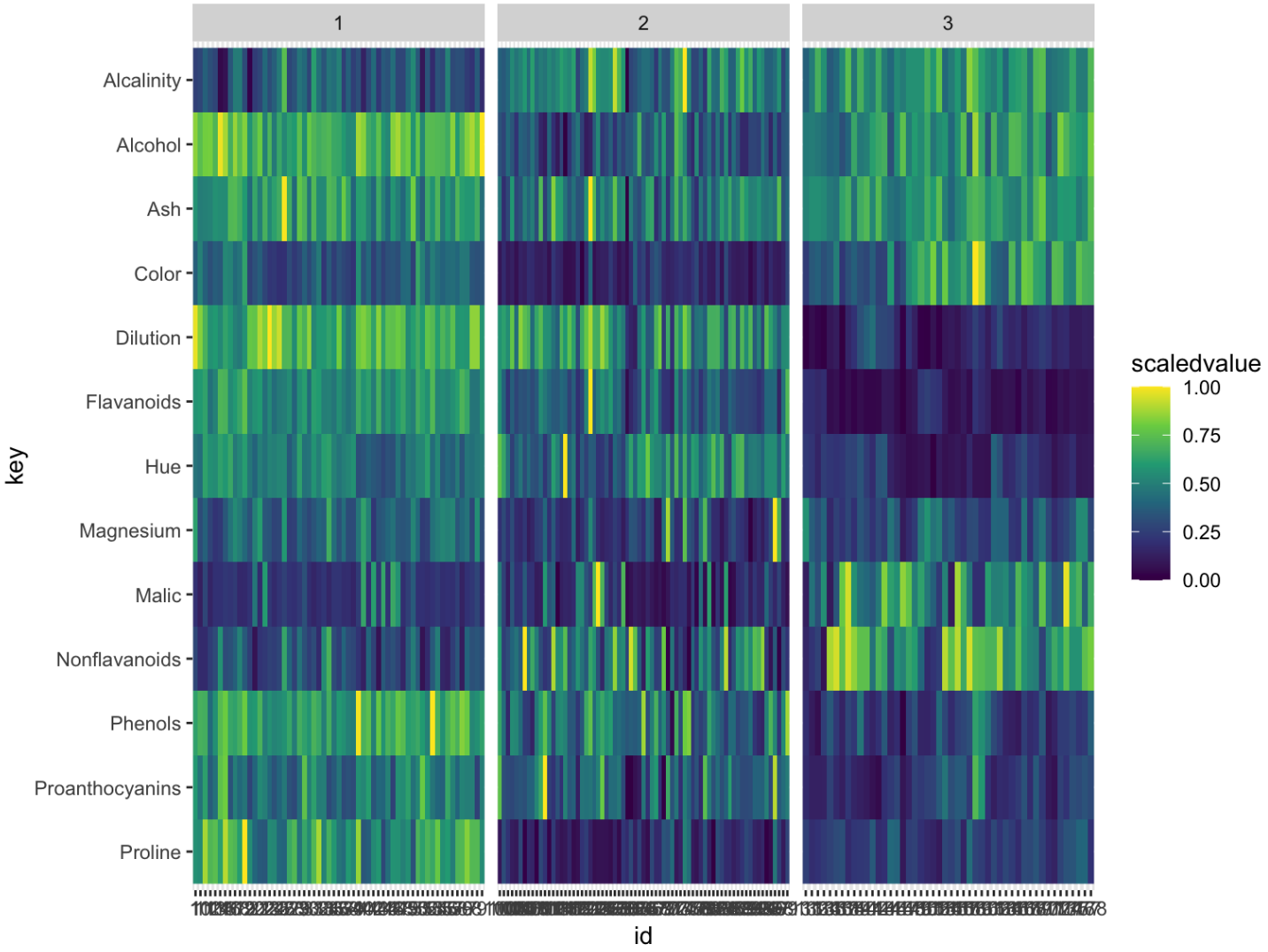


# Wine dataset




Alcalinity
Alcohol
Ash
Color
Dilution
Flavanoids
Hue
Magnesium
Malic
Nonflavanoids
Phenols
Proanthocyanins
Proline
key
id
scaledvalue
1.00
0.75
0.50
0.25
0.00

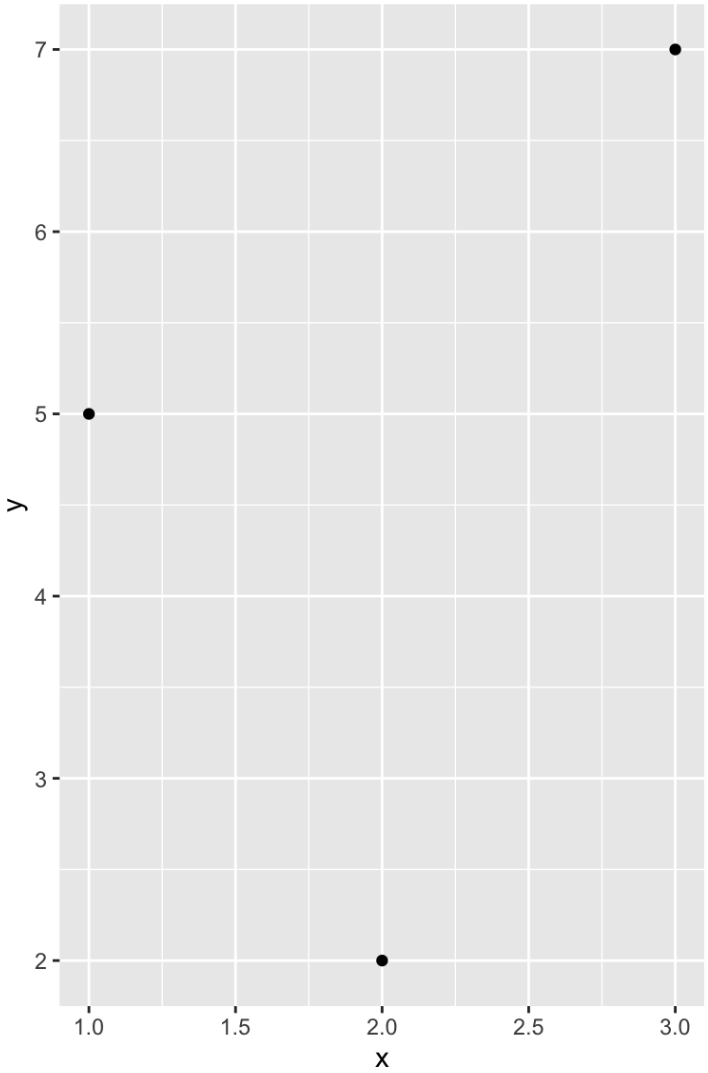
# Facet on type

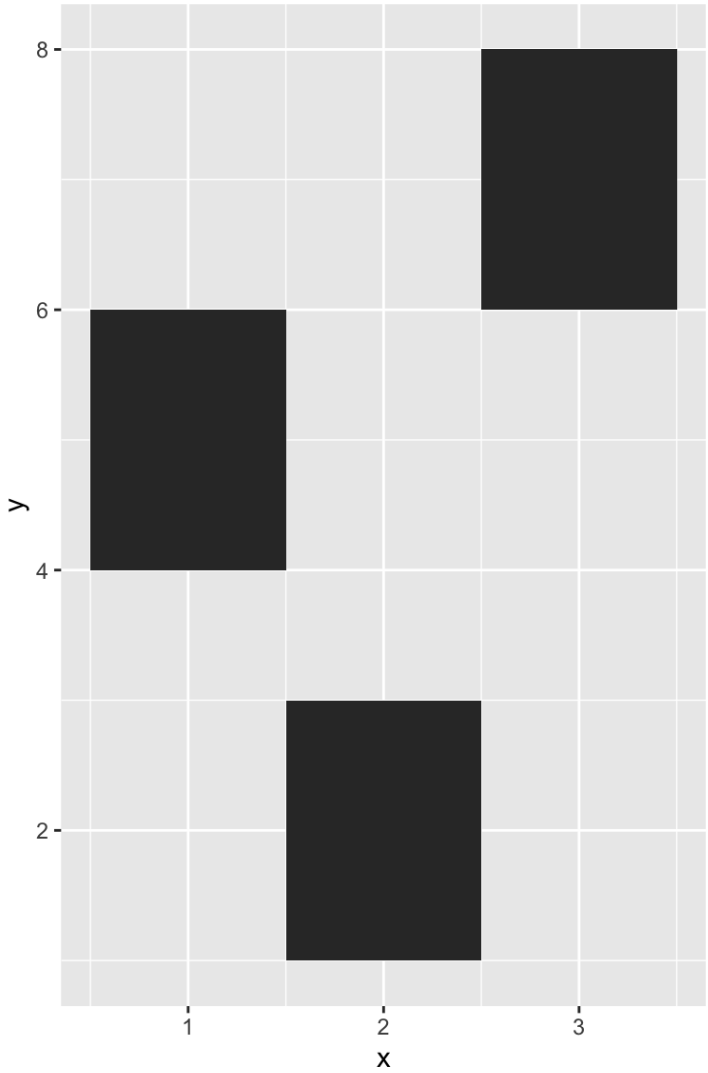


# Drawing heatmaps with `ggplot2`

`geom_tile` with numerical data, compare to `geom_point`

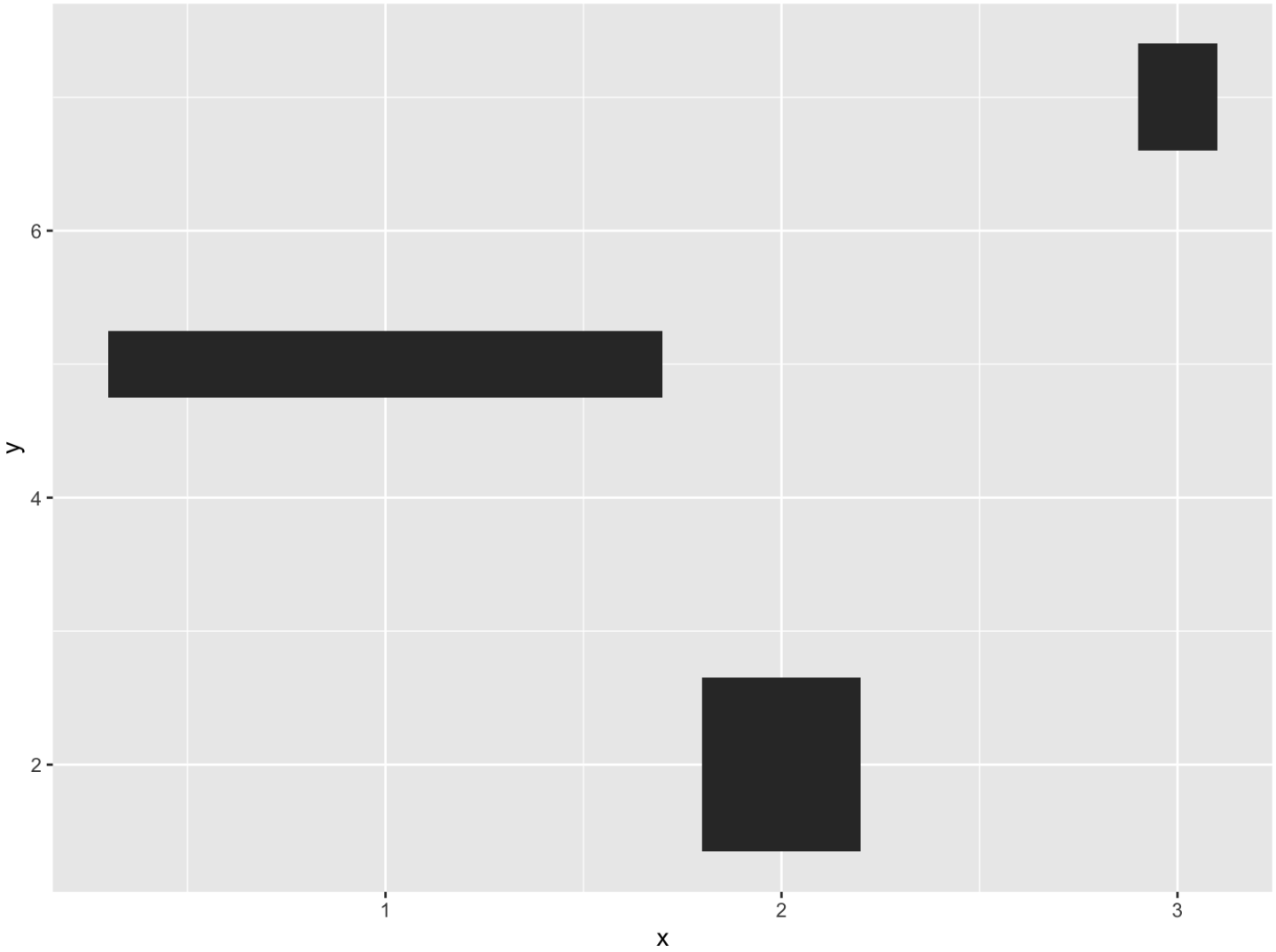
```
x <- 1:3
y <- c(5, 2, 7)
df <- data.frame(x, y)
g1 <- ggplot(df, aes(x, y)) + geom_point()
g2 <- ggplot(df, aes(x, y)) + geom_tile()
grid.arrange(g1, g2, nrow = 1)
```


7
6
5
4
3
2
y
1.0
1.5
2.0
2.5
3.0
x


8
6
4
2
y
1
2
3
x

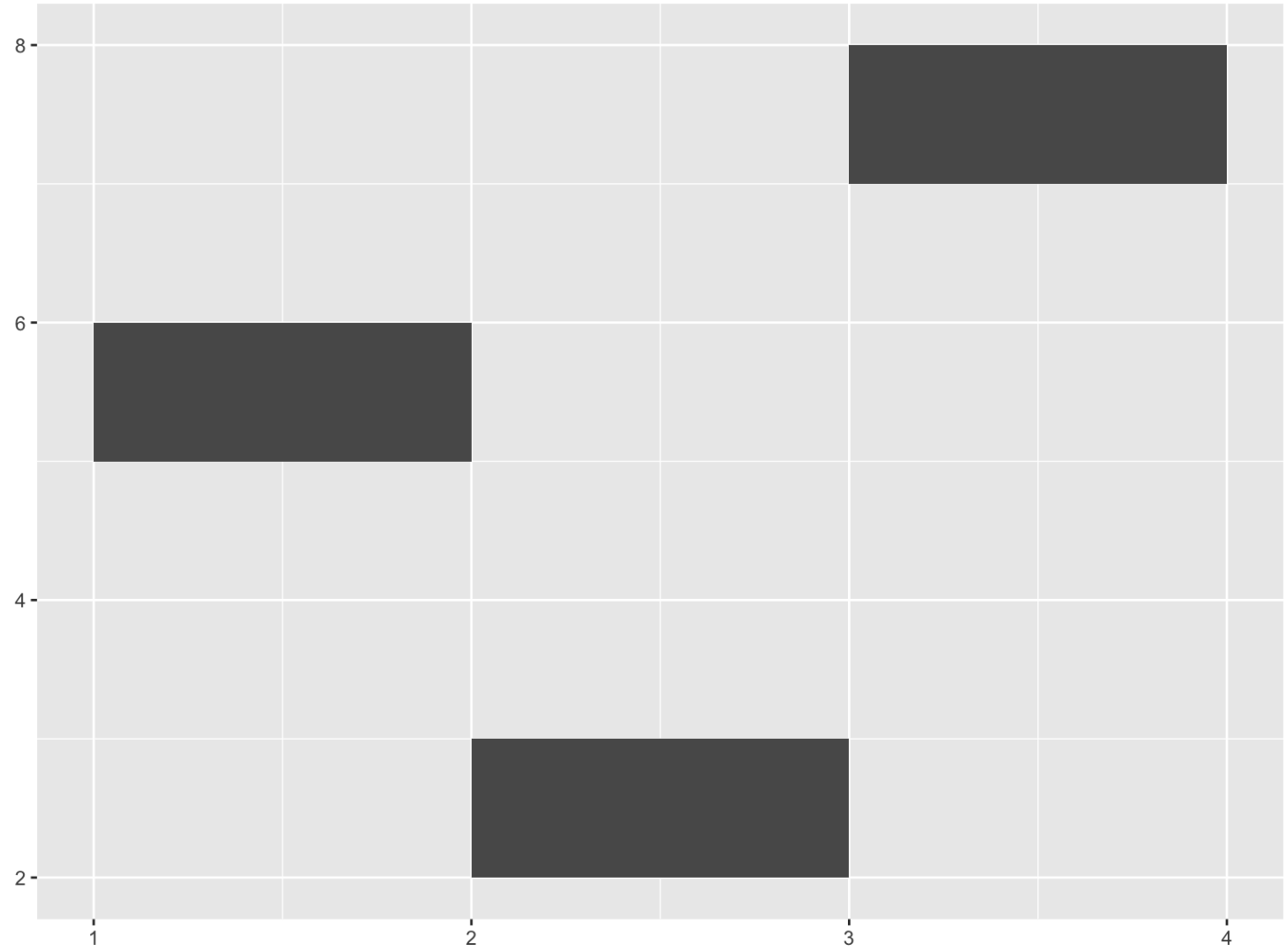
# geom_tile with variable w, h

```
df$w <- c(1.4, .4, .2)
df$h <- c(.5, 1.3, .8)
ggplot(df, aes(x, y)) + geom_tile(aes(width = w, height = h))
```



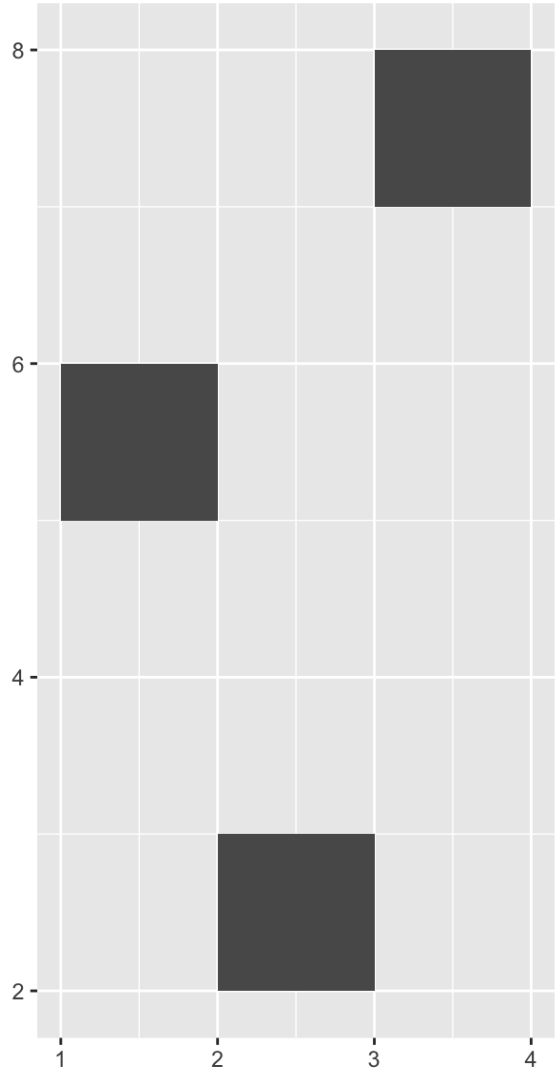
# geom_rect uses (xmin, xmax, ymin, ymax) instead

```
xmin <- 1:3
xmax <- 2:4
ymin <- c(5, 2, 7)
ymax <- c(6, 3, 8)
df <- data.frame(xmin, xmax, ymin, ymax)
ggplot(df, aes(xmin = xmin, xmax = xmax, ymin = ymin,
               ymax = ymax)) + geom_rect()
```



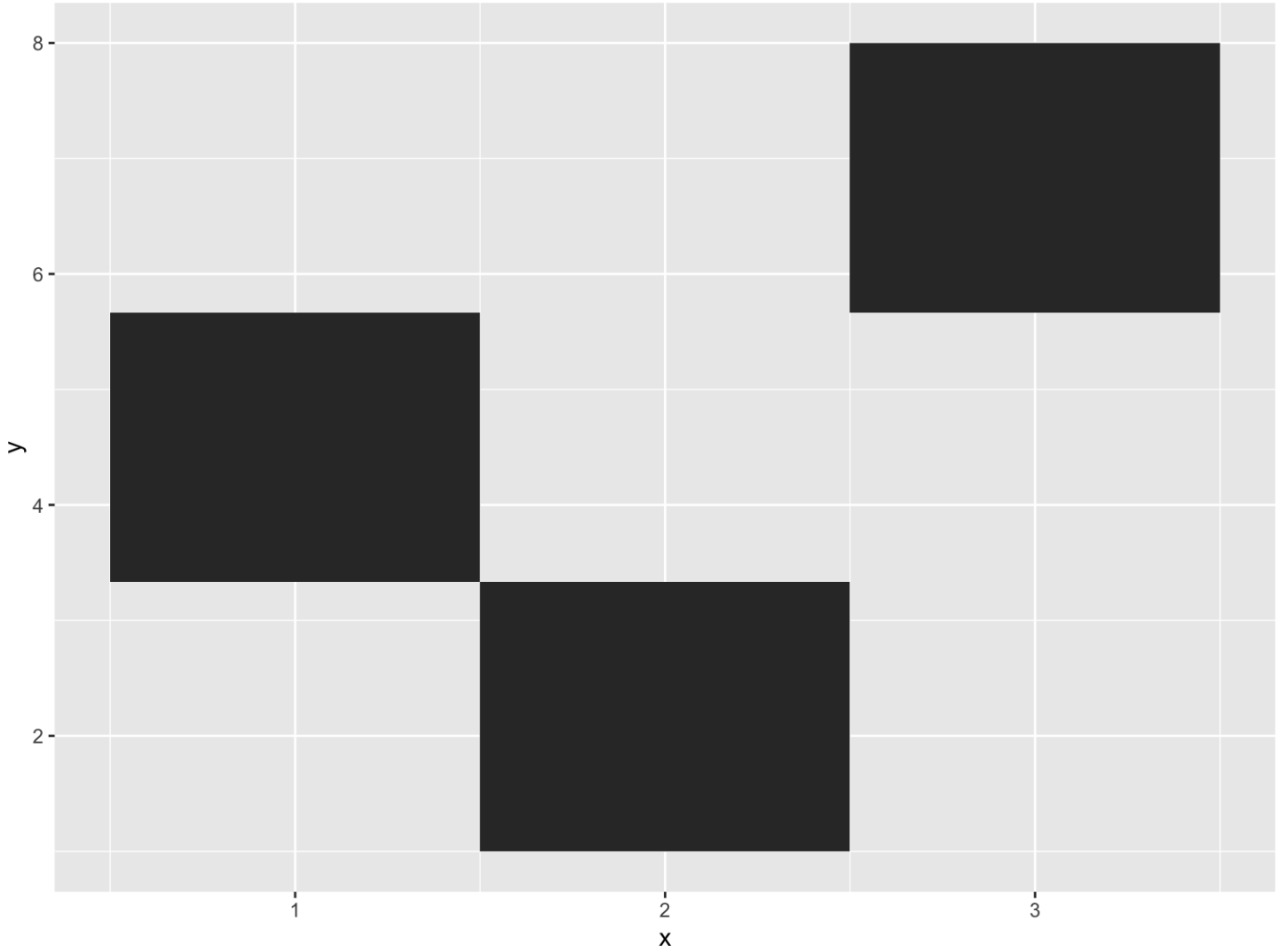
# force squares

```
ggplot(df, aes(xmin = xmin, xmax = xmax, ymin = ymin,
               ymax = ymax)) + geom_rect() + coord_fixed()
```



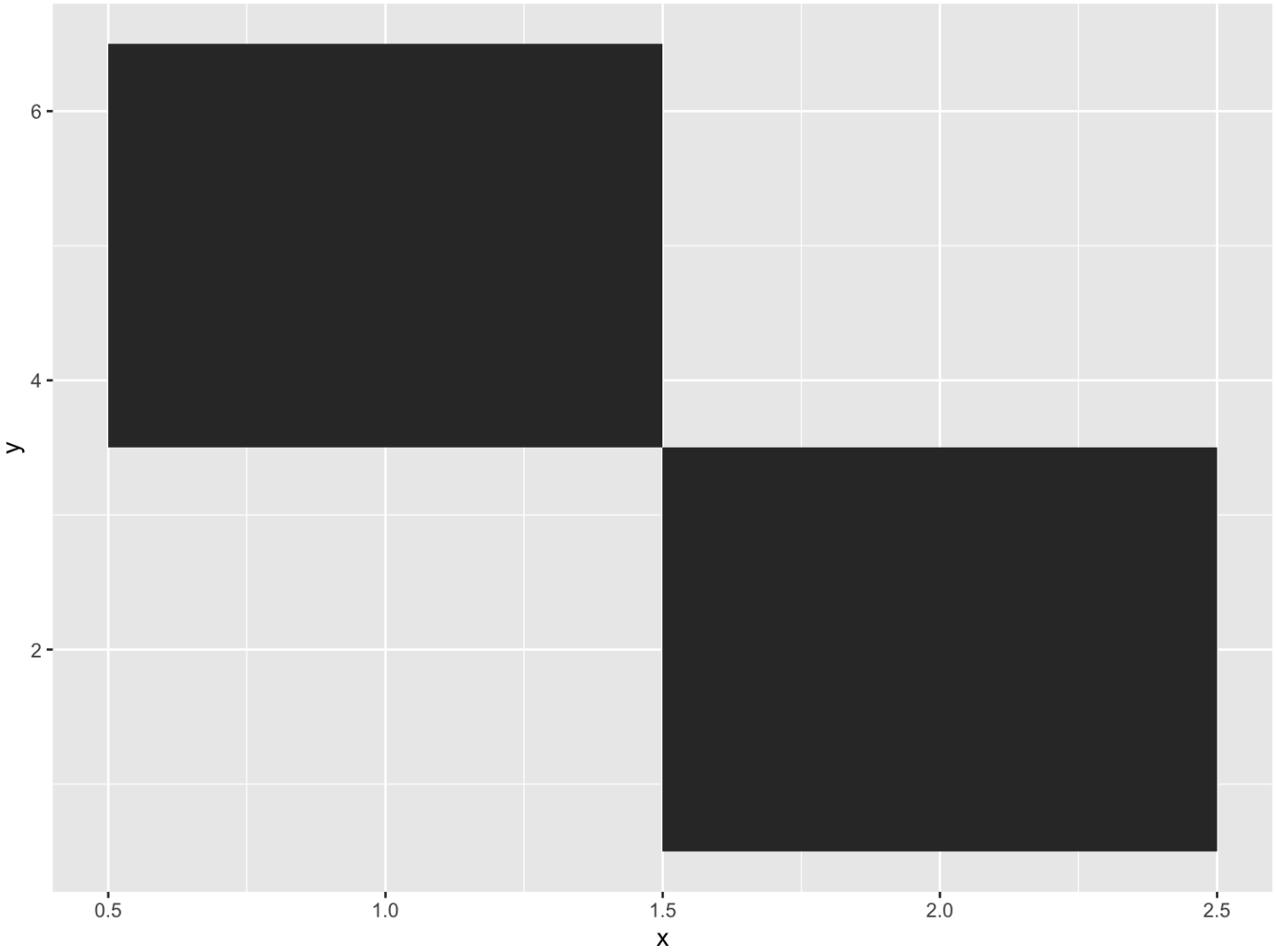
# geom_raster – same as geom_tile w/ uniform w, h & FASTER

```r
x <- 1:3
y <- c(5, 2, 7)
df <- data.frame(x, y)

ggplot(df, aes(x,y)) + geom_raster()
```



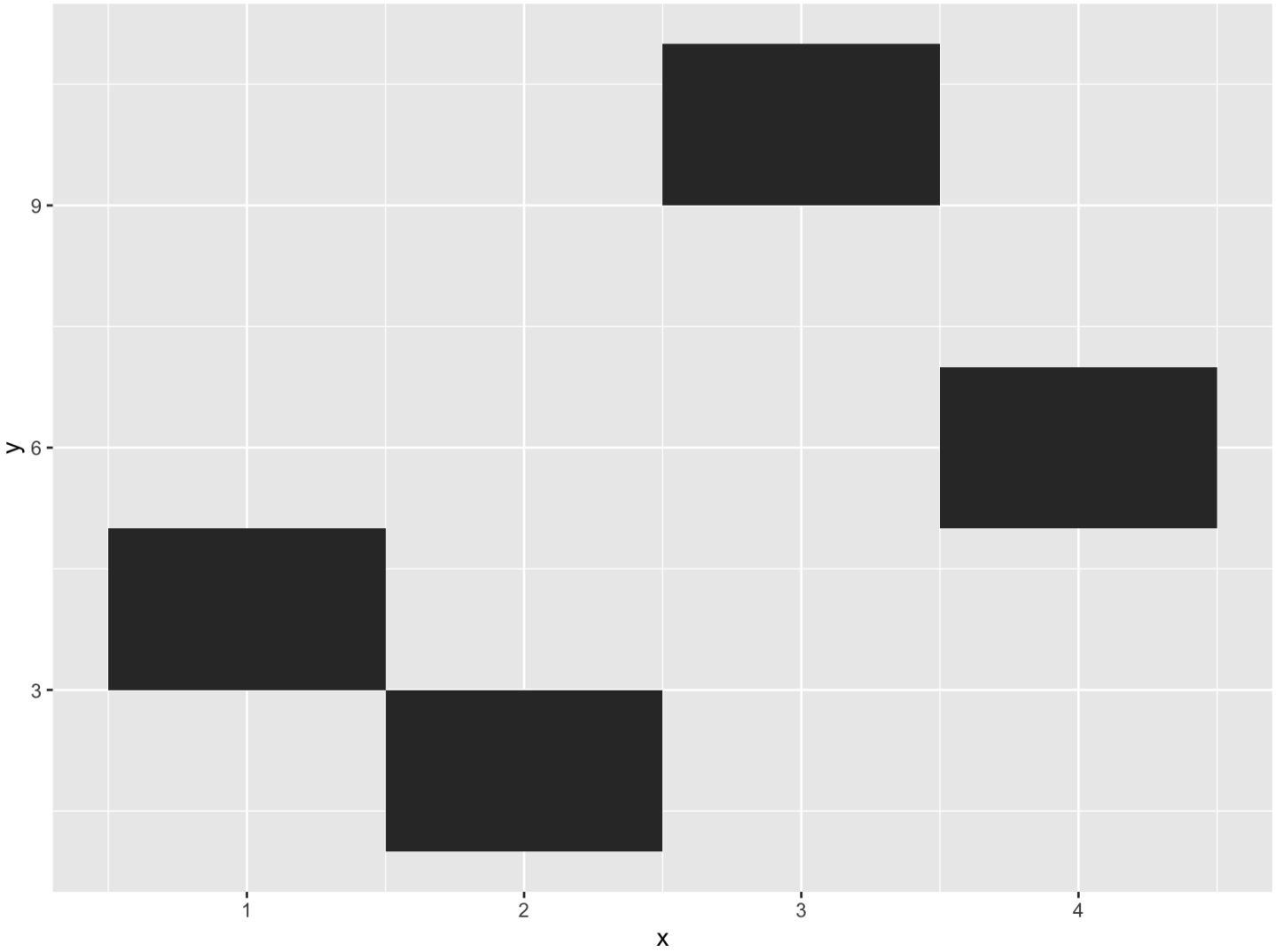
# change x & y: what happens to the size of the tiles?

```r
x <- 1:2
y <- c(5, 2)
df <- data.frame(x, y)
ggplot(df, aes(x, y)) + geom_raster()
```



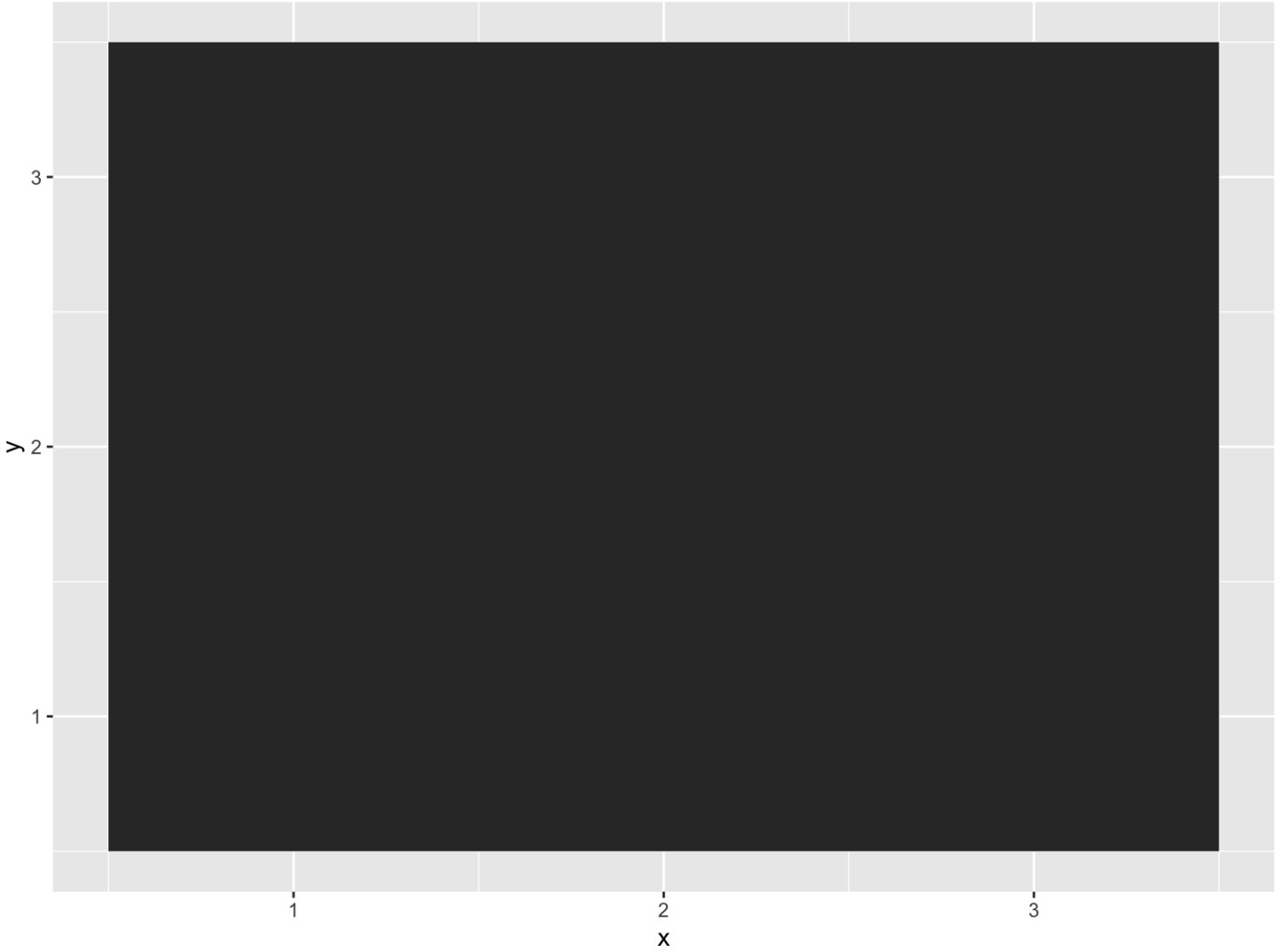
# change x & y again

```
x <- 1:4
y <- c(5, 2, 10, 7)
df <- data.frame(x, y)
ggplot(df, aes(x, y)) + geom_raster()
```



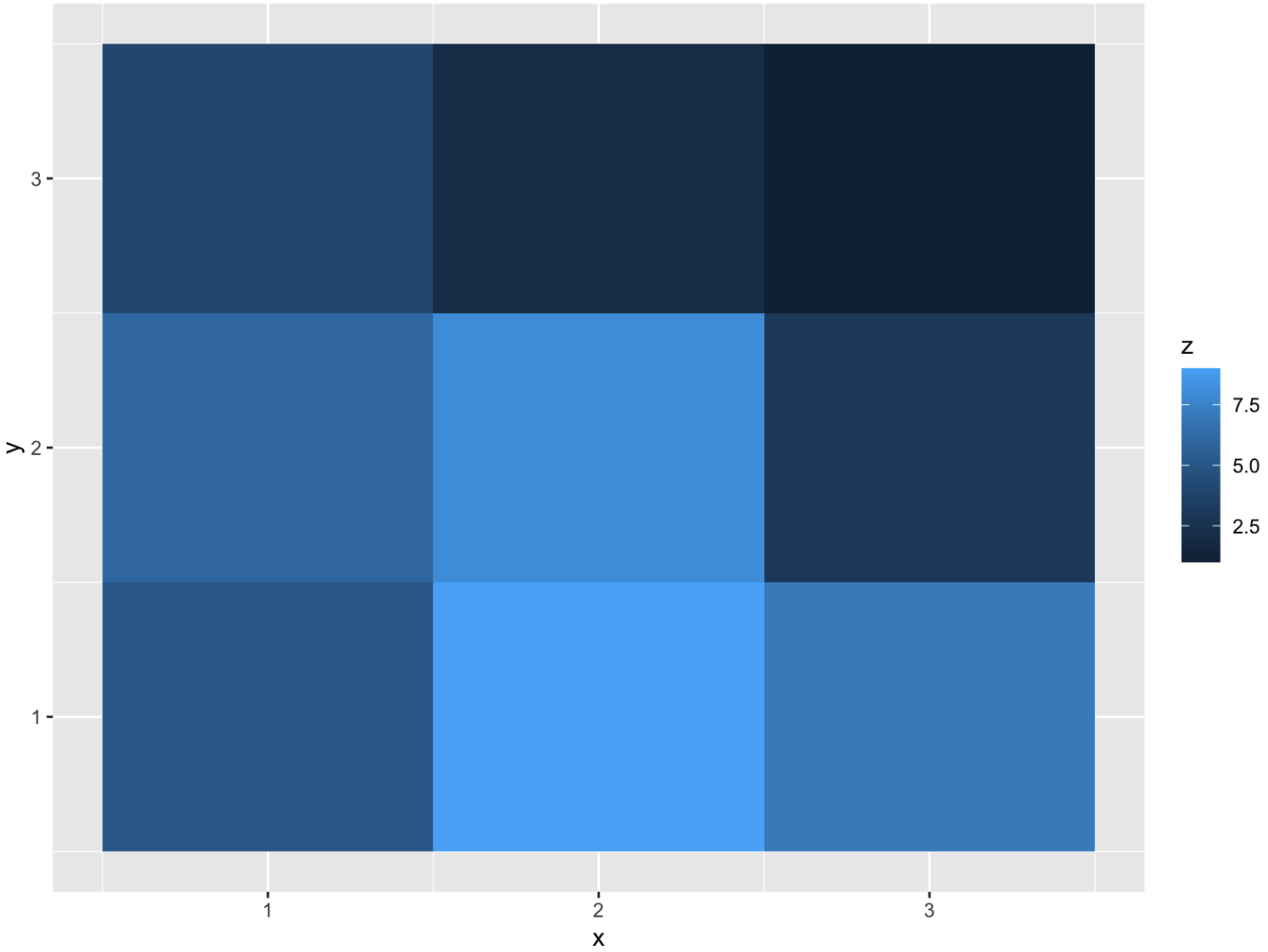
# complete set of (x, y) pairs

```
x <- c(1, 1, 1, 2, 2, 2, 3, 3, 3)
y <- c(1, 2, 3, 1, 2, 3, 1, 2, 3)
df <- data.frame(x, y)
ggplot(df, aes(x, y)) + geom_raster()
```



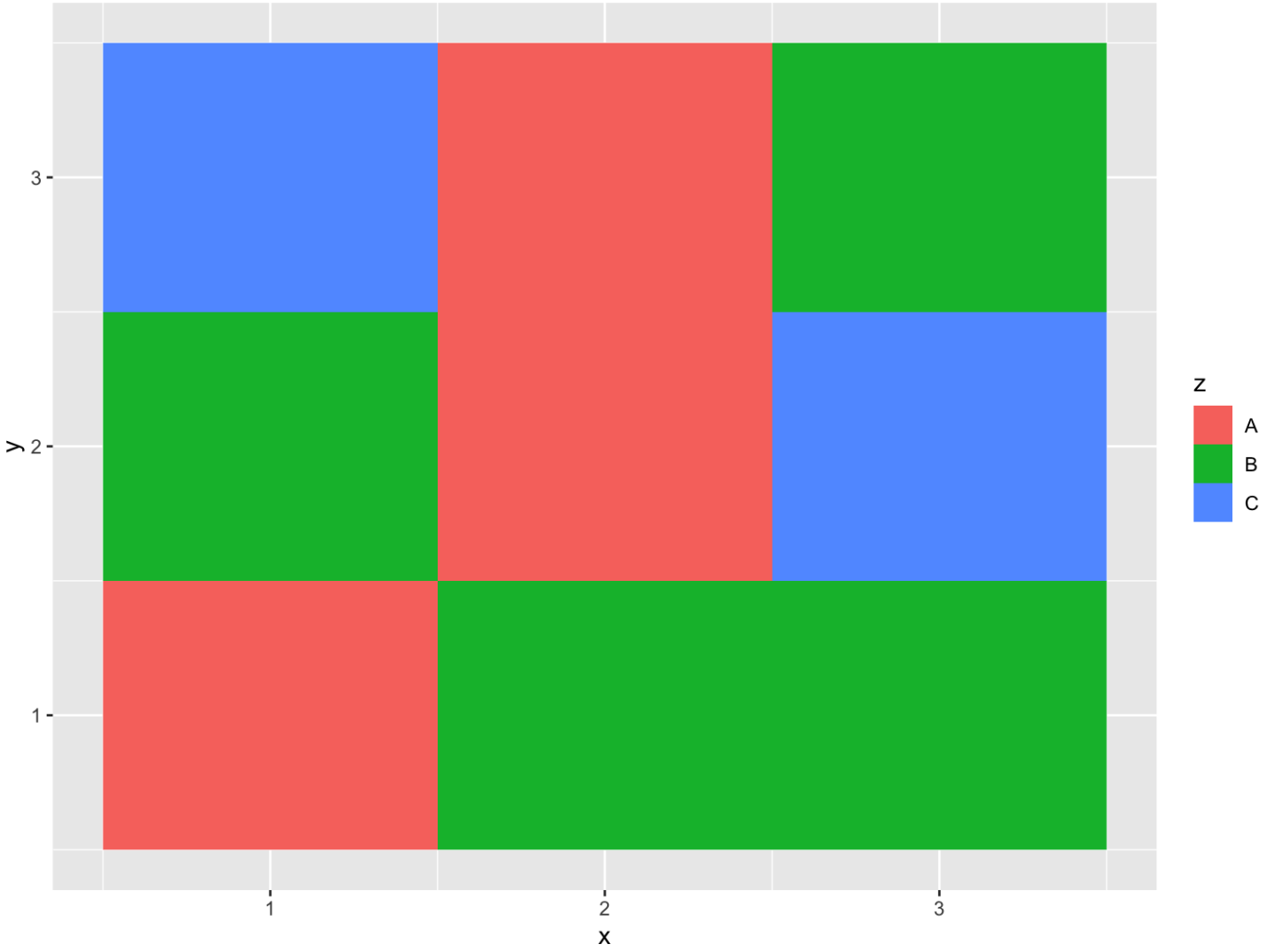
# add color

```
set.seed(2017)
df$z <- sample(9)
ggplot(df, aes(x, y)) + geom_raster(aes(fill = z))
```



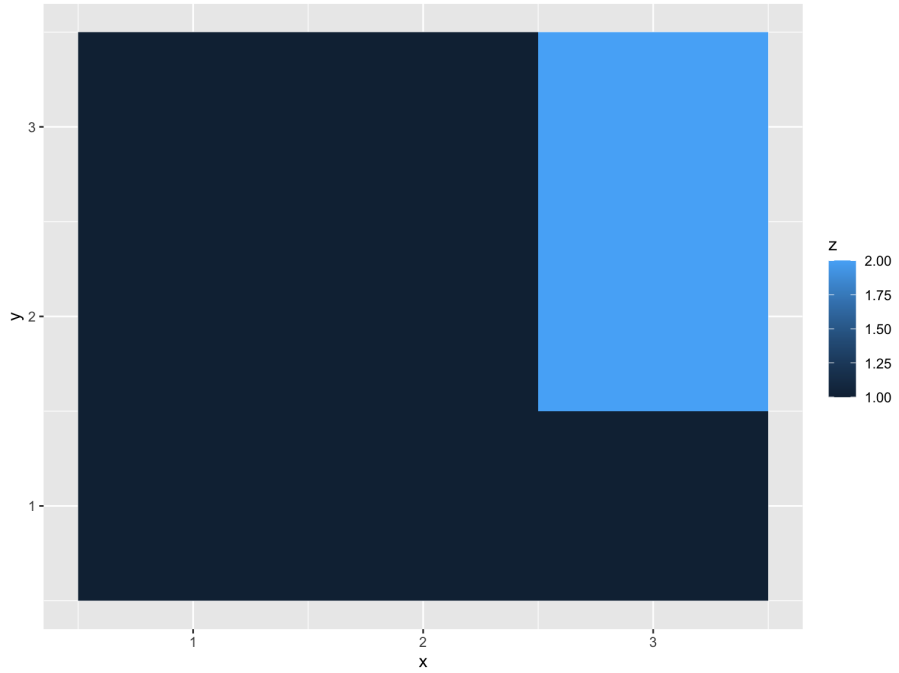
# What if z is categorical?

```
df$z <- c("A", "B", "C", "B", "A", "A", "B", "C", "B")
ggplot(df, aes(x, y)) + geom_raster(aes(fill = z))
```

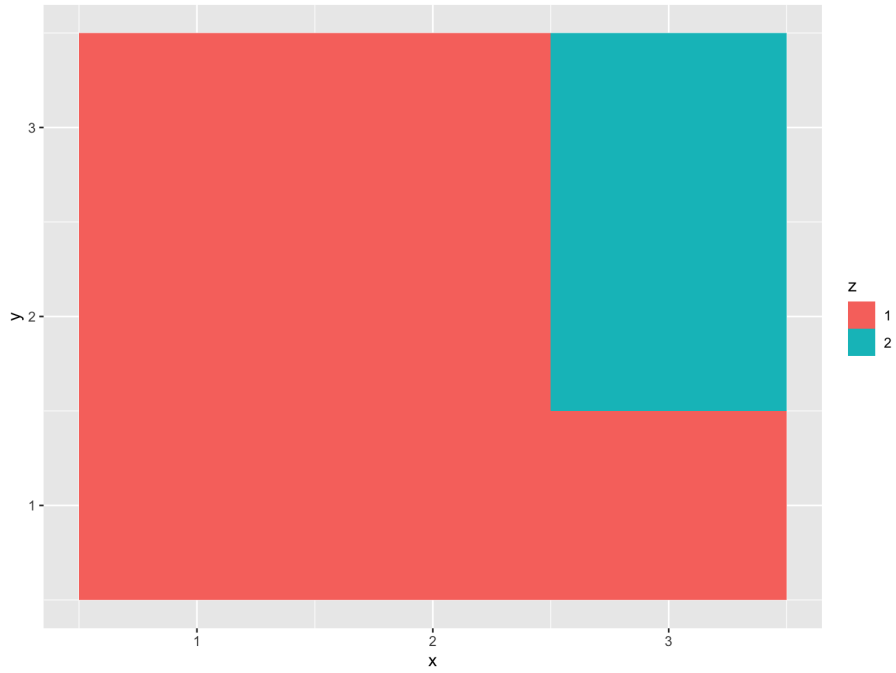


# What if z is discrete (numerical)?

```
df$z <- sample(3, 9, replace = TRUE)
ggplot(df, aes(x, y)) + geom_raster(aes(fill = z))
```

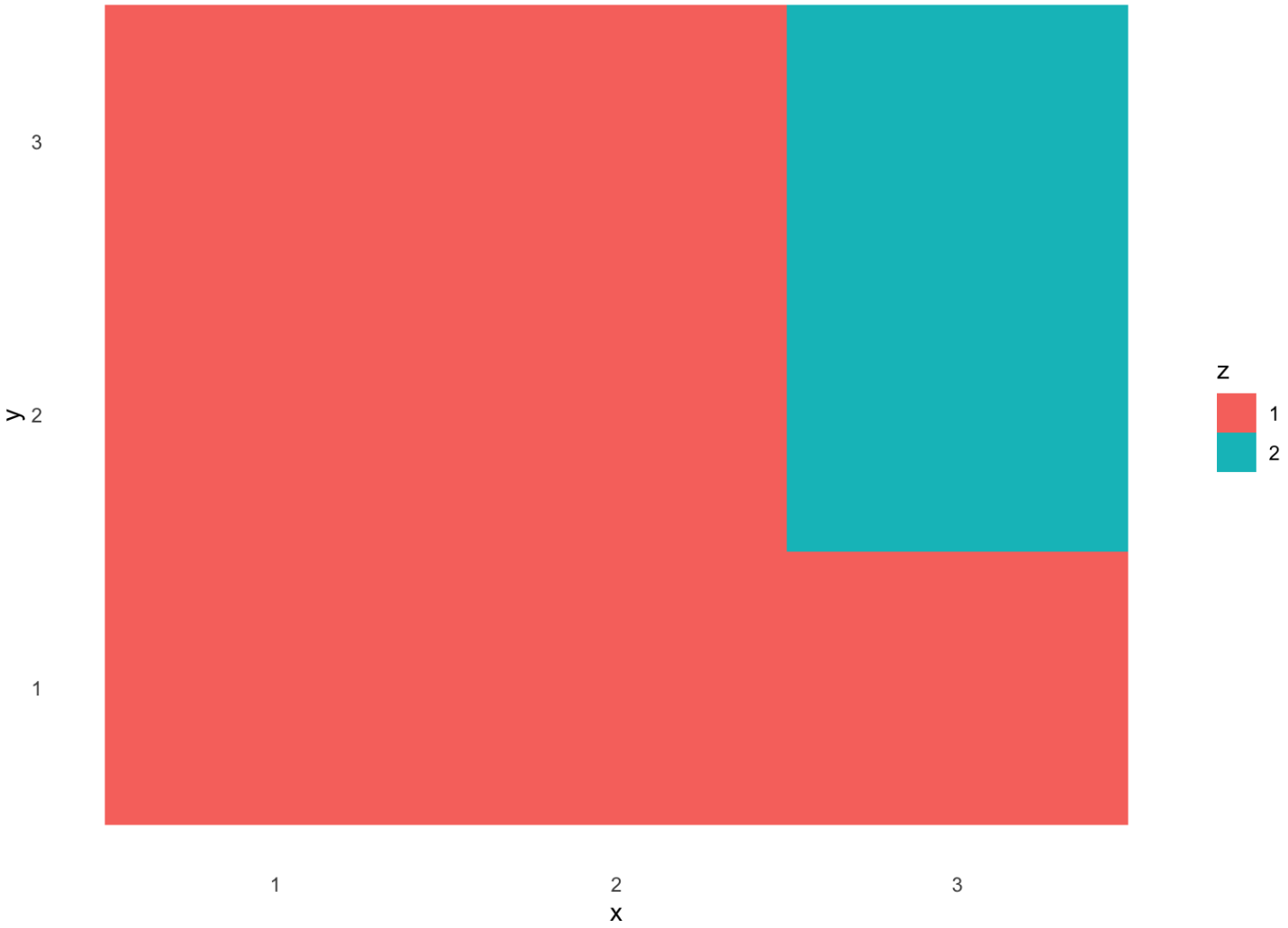



```
df$z <- factor(df$z)
ggplot(df, aes(x, y)) + geom_raster(aes(fill = z))
```


3
2
1
y
1
2
3
x
z
1
2

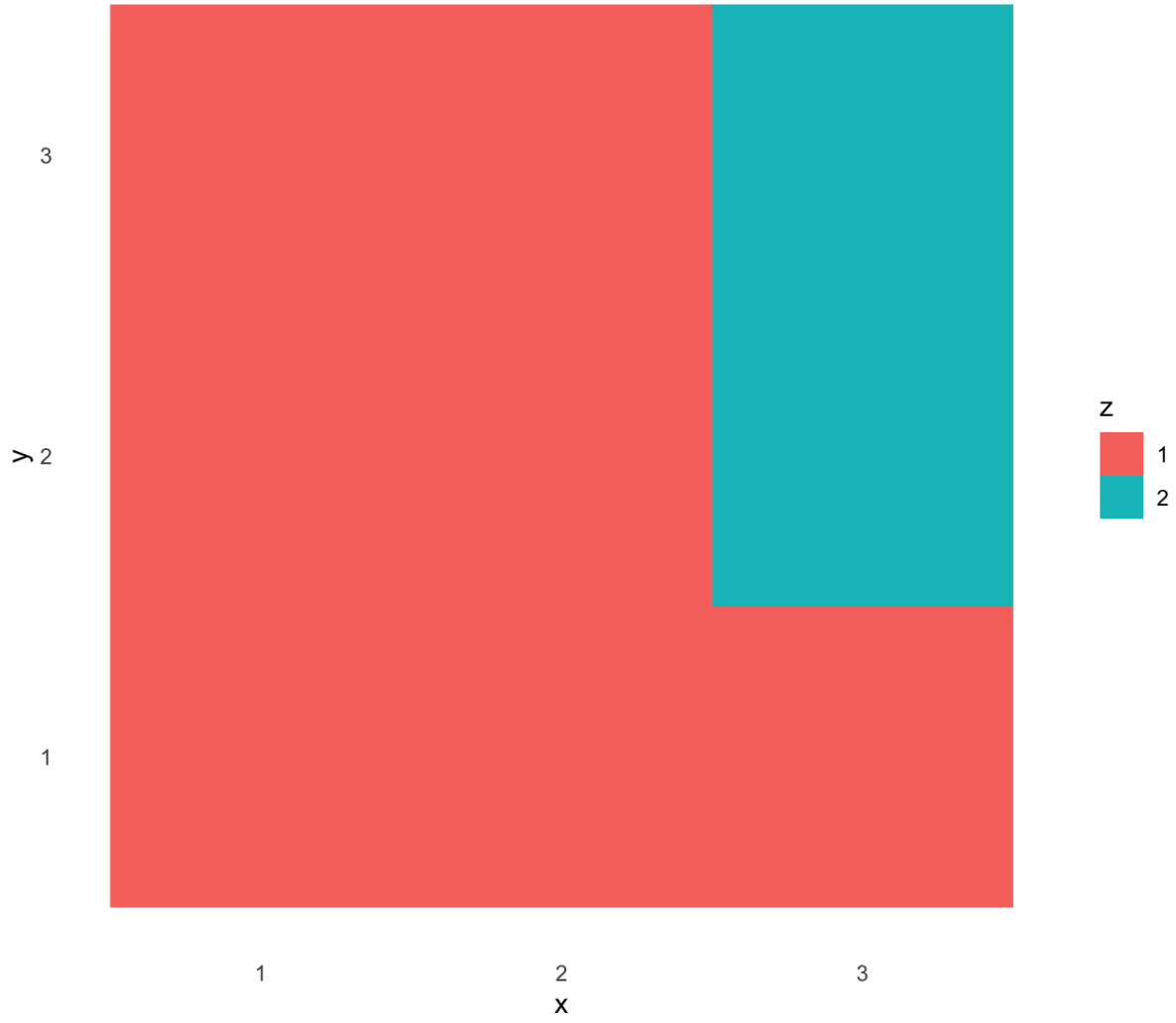
# create a heat map theme

```
theme_heat <- theme_classic() +
  theme(axis.line = element_blank(),
        axis.ticks = element_blank())

ggplot(df, aes(x, y)) + geom_raster(aes(fill = z)) +
  theme_heat
```



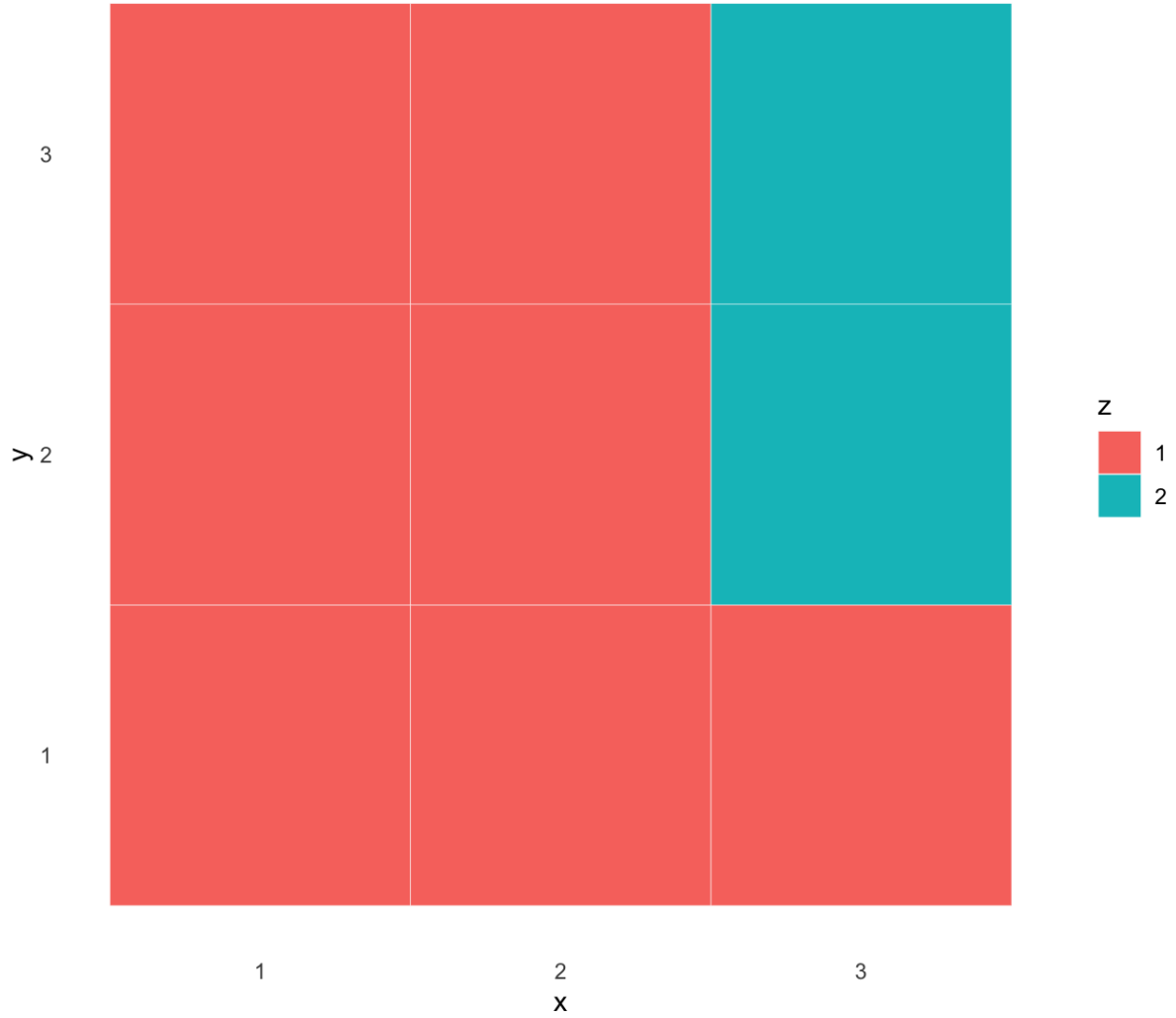
# add coord_fixed

```
ggplot(df, aes(x, y)) + geom_raster(aes(fill = z)) +
  coord_fixed() + theme_heat
```



# add white border

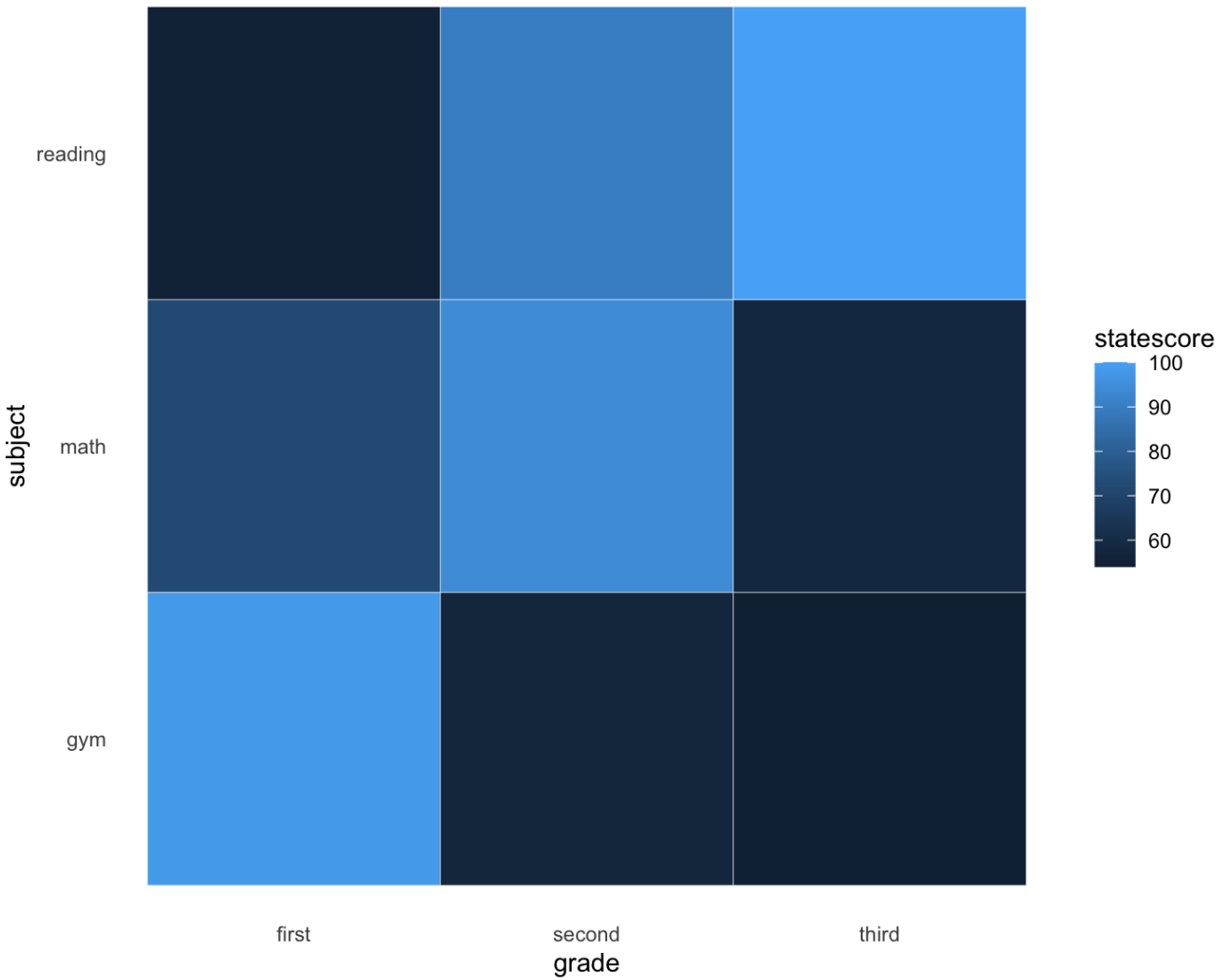
```
ggplot(df, aes(x, y)) +
  geom_tile(aes(fill = z), color = "white") +
  coord_fixed() + theme_heat
```




(doesn't work with `geom_raster()`)

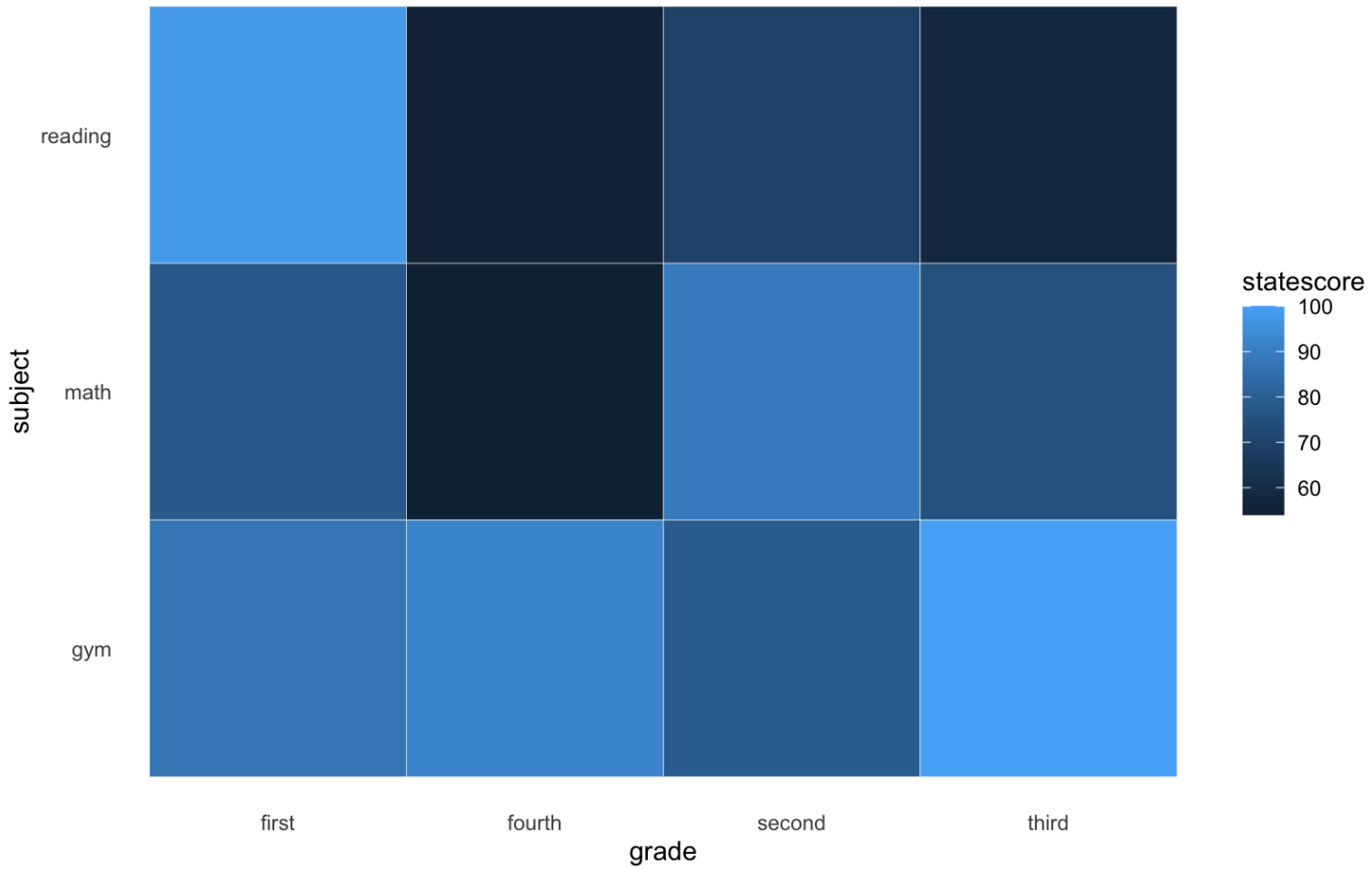
# x & y are categorical

```
grade <- rep(c("first", "second", "third"), 3)
subject <- rep(c("math", "reading", "gym"), each = 3)
statescore <- sample(50, 9) + 50
df <- data.frame(grade, subject, statescore)

ggplot(df, aes(grade, subject, fill = statescore)) +
  geom_tile(color = "white") +
  coord_equal() + theme_heat
```


reading
math
gym
subject
first
second
third
grade
statescore
100
90
80
70
60

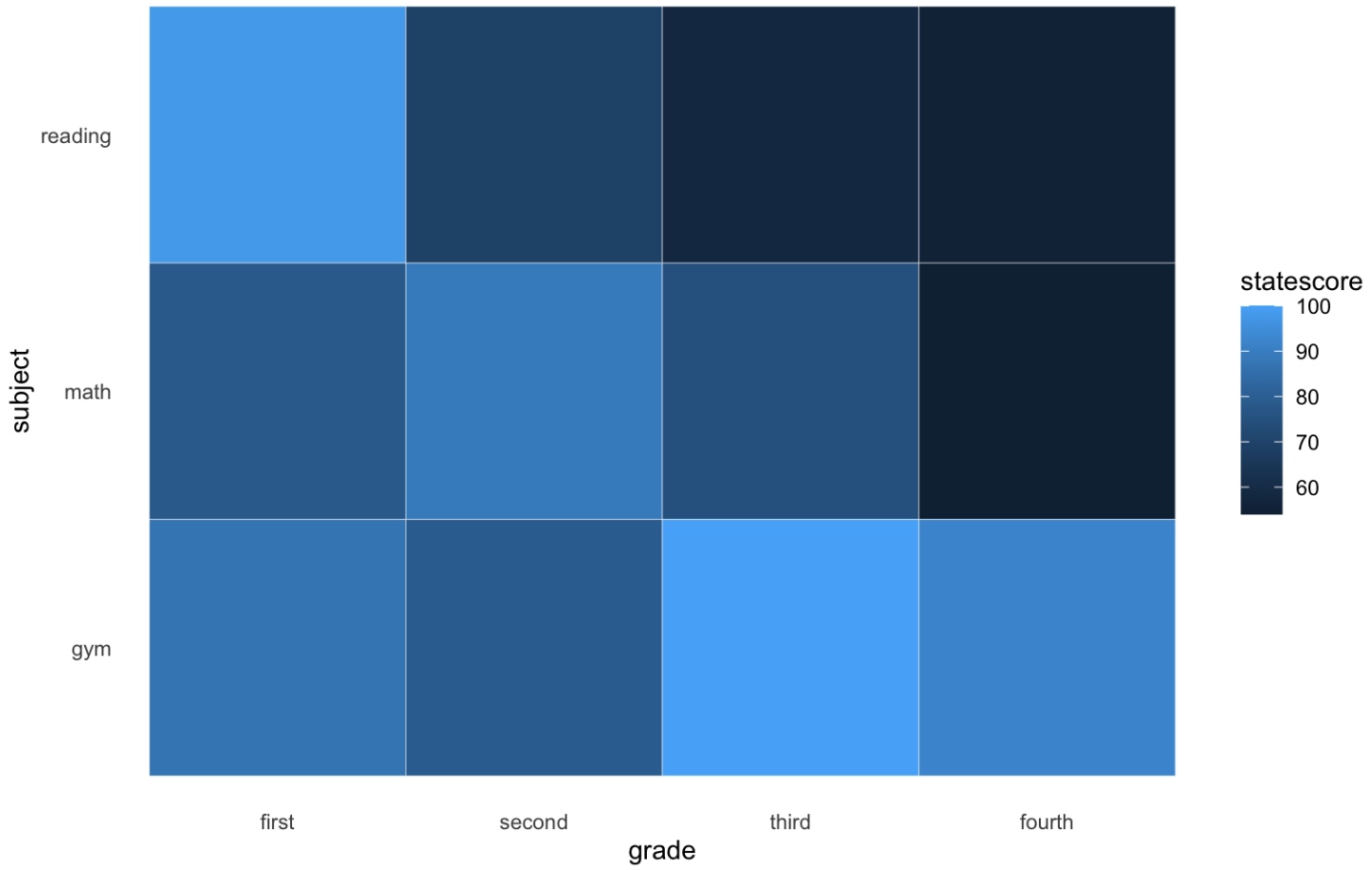
# problem with order of categories

```
grade <- rep(c("first", "second", "third", "fourth"), 3)
subject <- rep(c("math", "reading", "gym"), each = 4)
statescore <- sample(50, 12) + 50
df <- data.frame(grade, subject, statescore)

ggplot(df, aes(grade, subject, fill = statescore)) +
  geom_tile(color = "white") +
  coord_equal() + theme_heat
```


reading
math
gym
subject
first
fourth
second
third
grade
statescore
100
90
80
70
60

# fix order

```
df$grade <- forcats::fct_relevel(df$grade, "fourth", after = Inf)

ggplot(df, aes(grade, subject, fill = statescore)) +
  geom_tile(color = "white") +
  coord_equal() + theme_heat
```



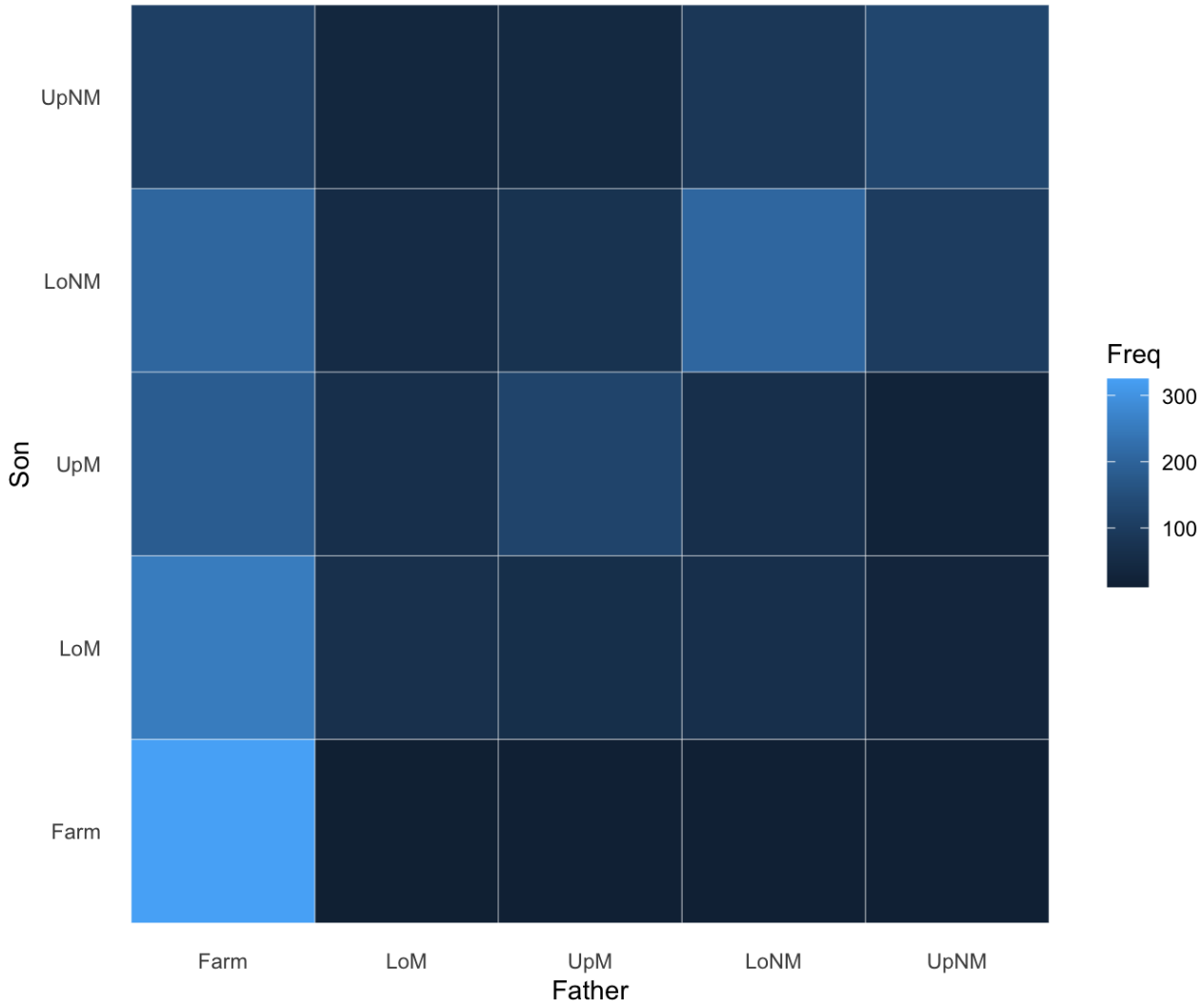
# Compare Japan, the UK, and the US in terms of occupational mobility.

```
library(vcdExtra)
library(dplyr)
orderedclasses <- c("Farm", "LoM", "UpM", "LoNM", "UpNM")
mydata <- Yamaguchi87
mydata$Son <- factor(mydata$Son, levels = orderedclasses)
mydata$Father <- factor(mydata$Father,
                        levels = orderedclasses)
japan <- mydata %>% filter(Country == "Japan")
uk <- mydata %>% filter(Country == "UK")
us <- mydata %>% filter(Country == "US")
```

# Compare Japan, the UK, and the US in terms of occupational mobility.

## Japan

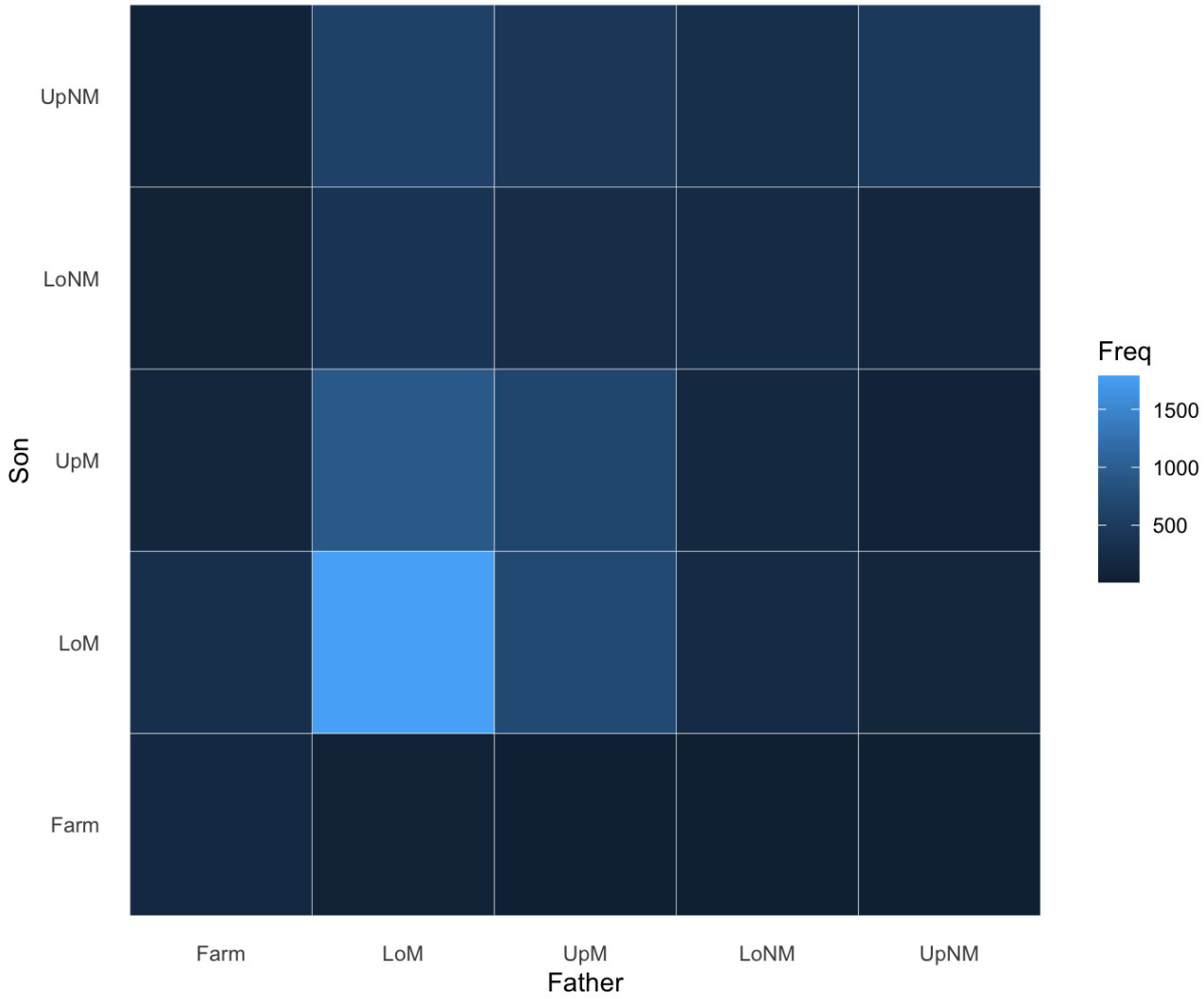
```
ggplot(japan, aes(x = Father, y = Son)) +
  geom_tile(aes(fill = Freq), color = "white") +
  coord_fixed() + theme_heat
```


UpNM
LoNM
UpM
LoM
Farm
Son
Farm
LoM
UpM
LoNM
UpNM
Father
Freq
300
200
100

# Compare Japan, the UK, and the US in terms of occupational mobility.

## UK

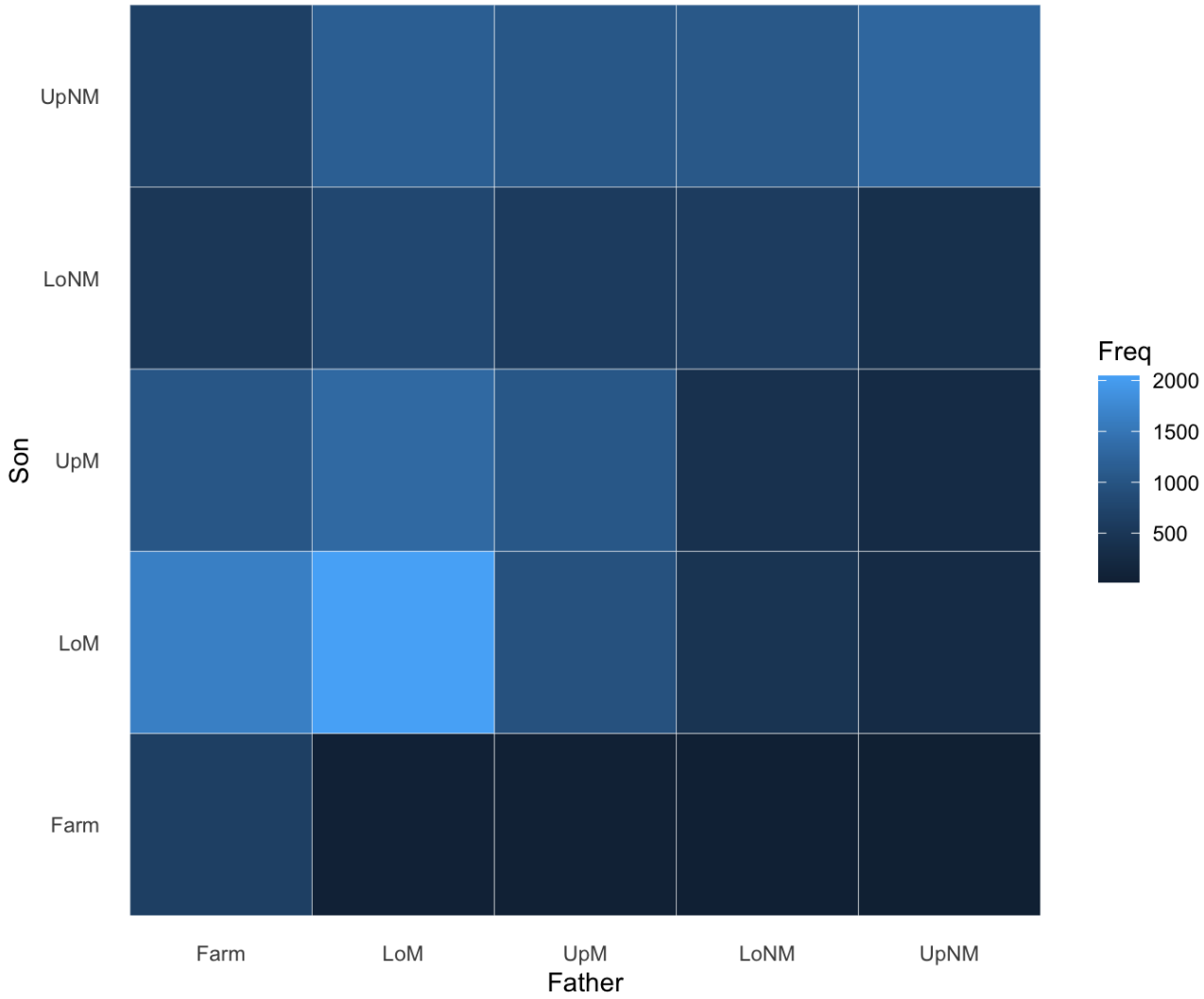
```
ggplot(uk, aes(x = Father, y = Son)) +
  geom_tile(aes(fill = Freq), color = "white") +
  coord_fixed() + theme_heat
```


UpNM
LoNM
UpM
LoM
Farm
Son
Farm
LoM
UpM
LoNM
UpNM
Father
Freq
1500
1000
500

# Compare Japan, the UK, and the US in terms of occupational mobility.

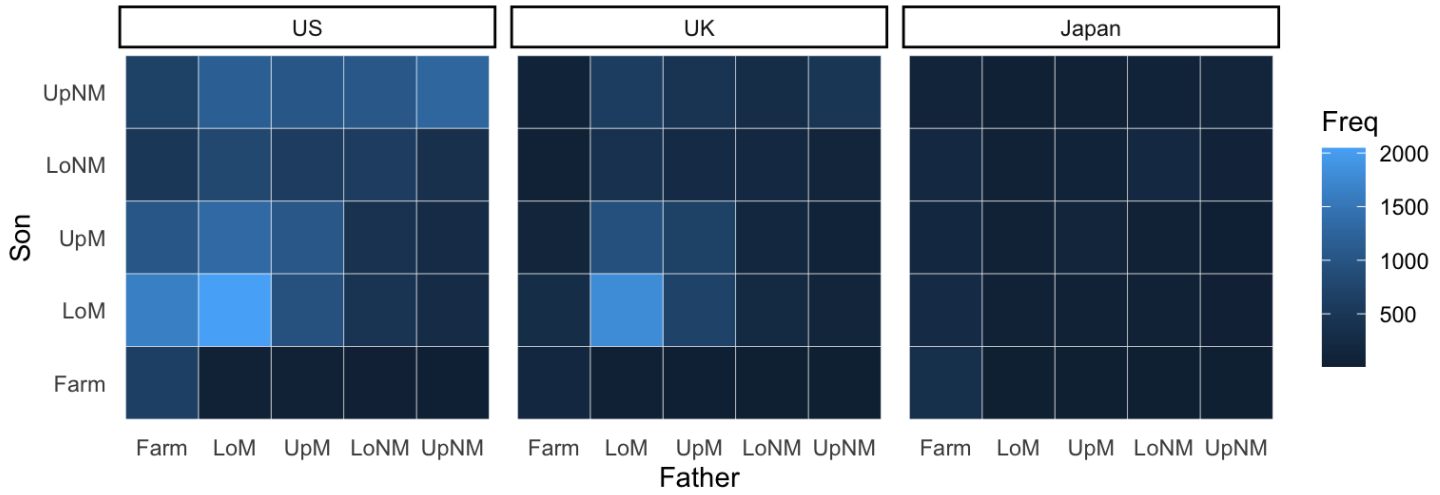
## US

```
ggplot(us, aes(x = Father, y = Son)) +
  geom_tile(aes(fill = Freq), color = "white") +
  coord_fixed() + theme_heat
```


UpNM
LoNM
UpM
LoM
Farm
Son
Farm
LoM
UpM
LoNM
UpNM
Father
Freq
2000
1500
1000
500

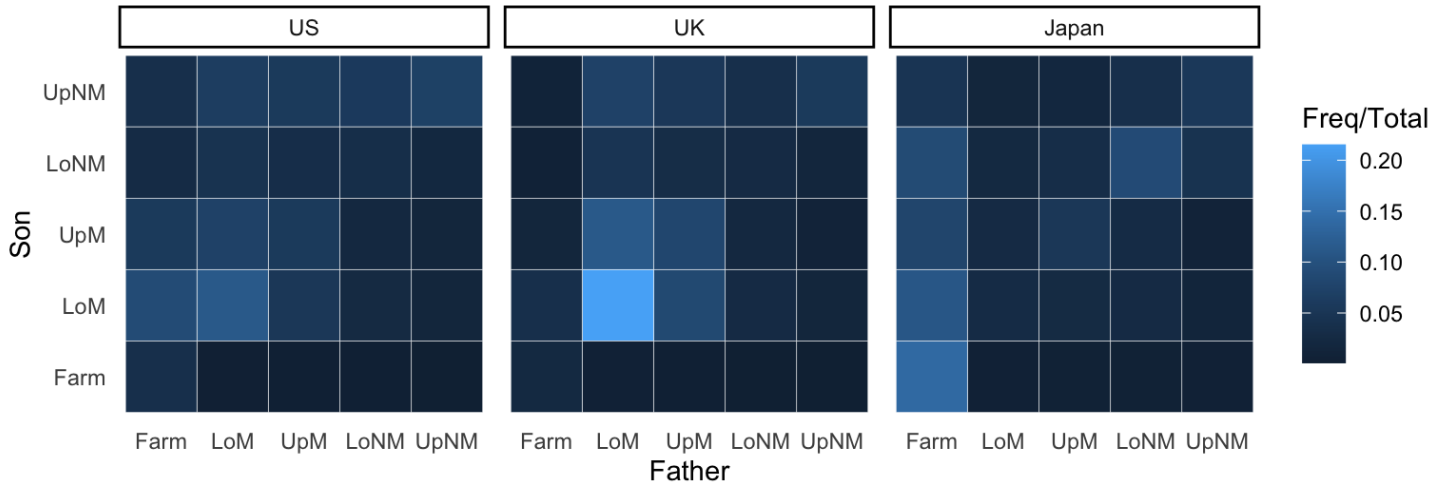
# All

```
ggplot(mydata, aes(x = Father, y = Son)) +
  geom_tile(aes(fill = Freq), color = "white") +
  coord_fixed() + facet_wrap(~Country) + theme_heat
```



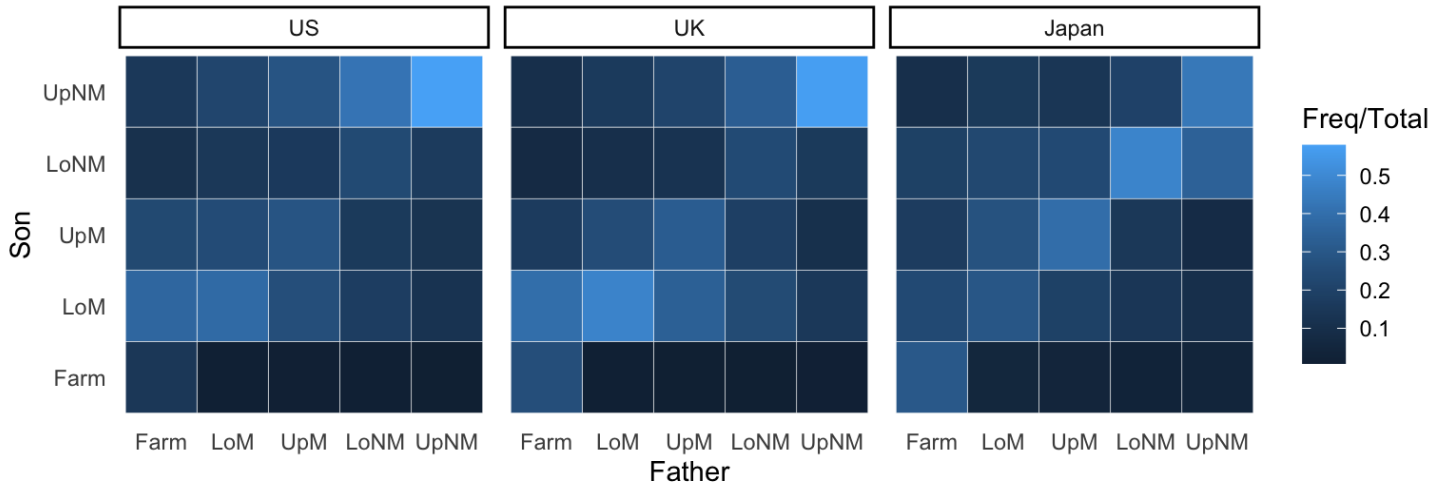
# All, as % of country total

```
mydata2 <- mydata %>% group_by(Country) %>%
  mutate(Total = sum(Freq)) %>% ungroup()

ggplot(mydata2, aes(x = Father, y = Son)) +
  geom_tile(aes(fill = Freq/Total), color = "white") +
  coord_fixed() + facet_wrap(~Country) + theme_heat
```

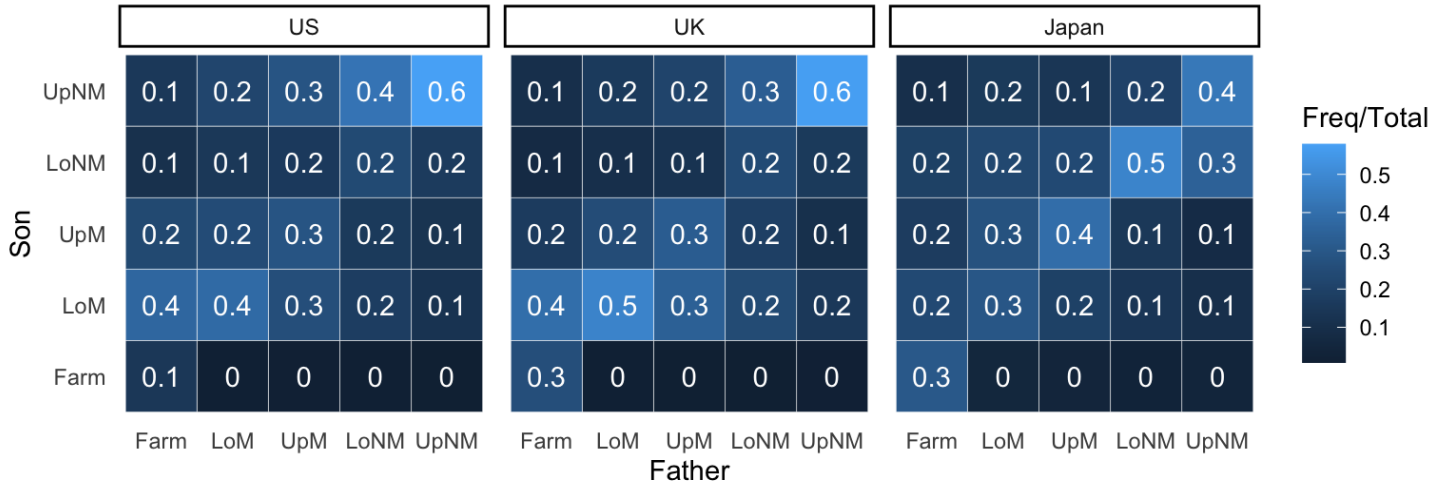


# All, as % of country and class total

```
mydata3 <- mydata %>% group_by(Country, Father) %>%
  mutate(Total = sum(Freq)) %>% ungroup()

g <- ggplot(mydata3, aes(x = Father, y = Son)) +
  geom_tile(aes(fill = Freq/Total), color = "white") +
  coord_fixed() + facet_wrap(~Country) + theme_heat
g
```

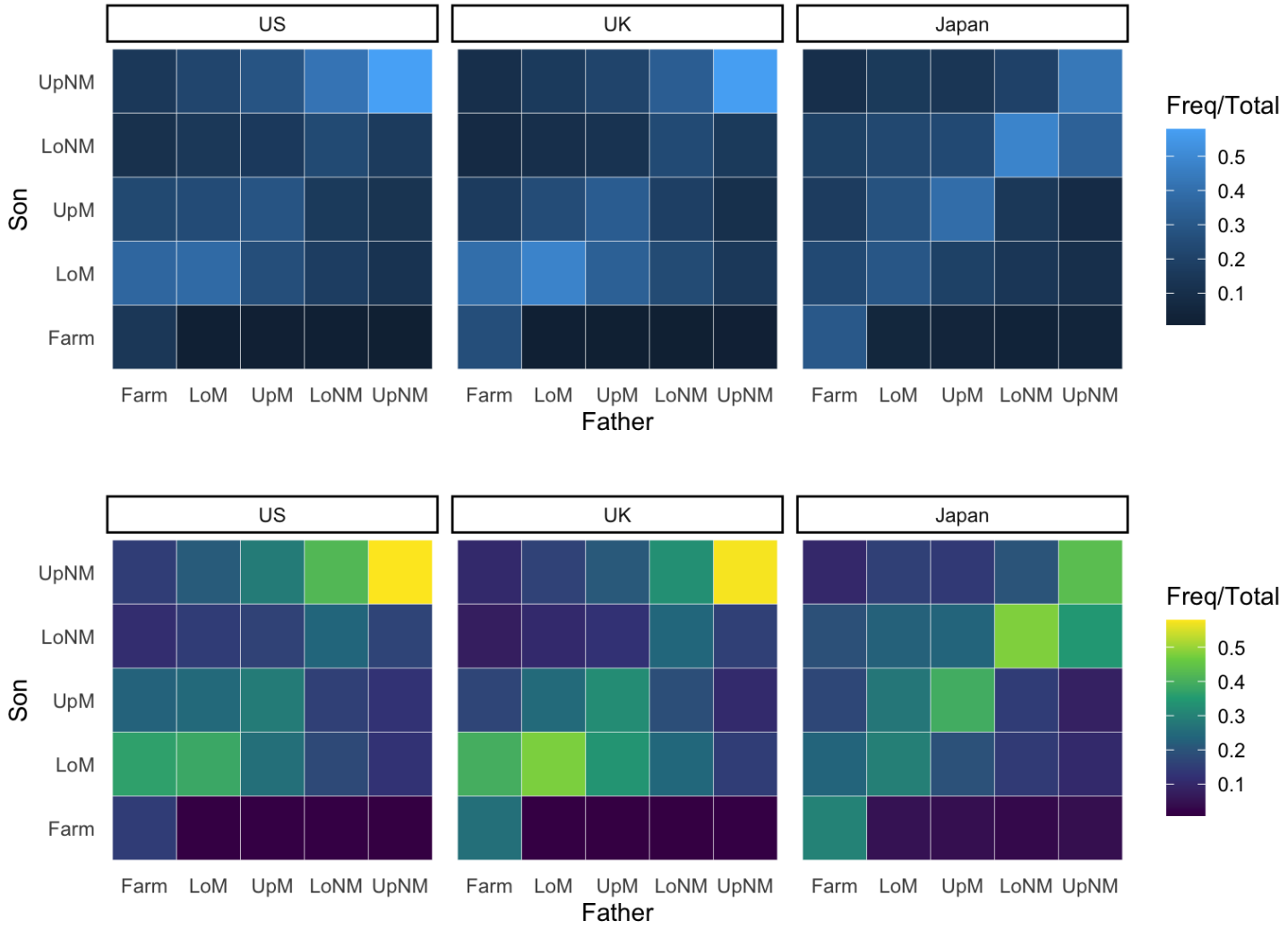


```
g + geom_text(aes(label = round(Freq/Total, 1)),
              color = "white")
```

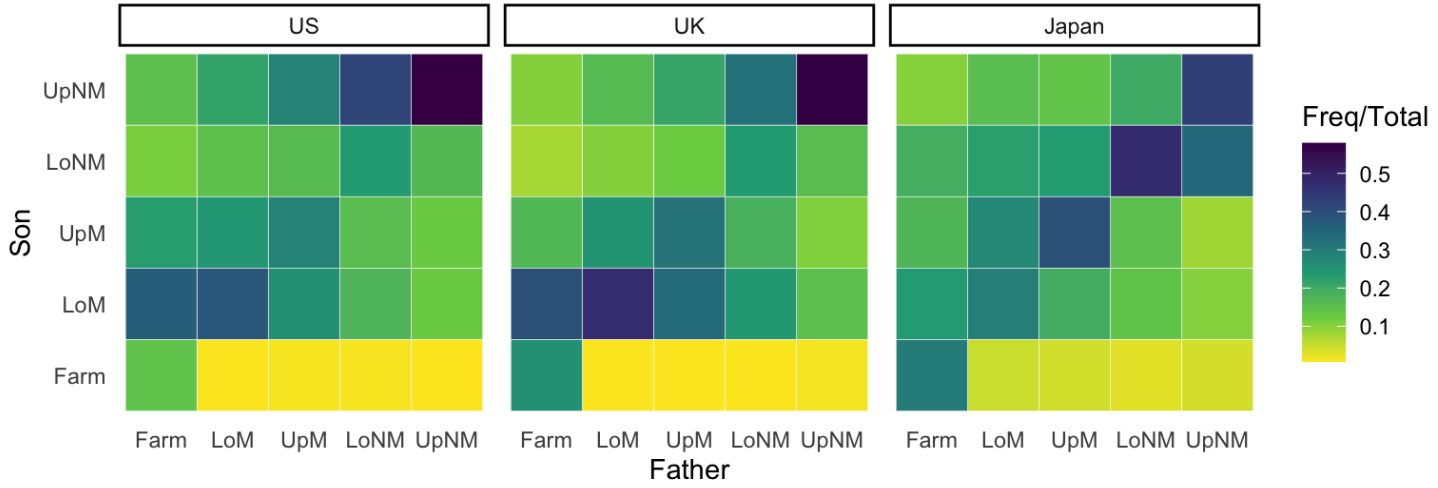


# Yamaguchi87

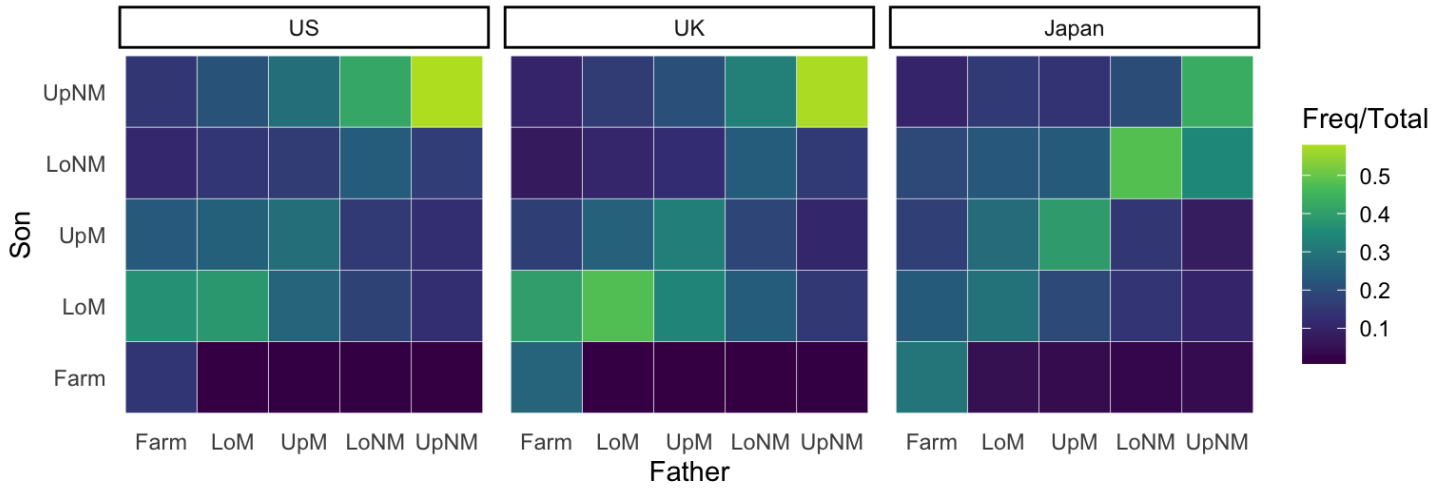
```
grid.arrange(g, g + scale_fill_viridis_c(),
             nrow = 2)
```



```
g + scale_fill_viridis_c(direction = -1)
```

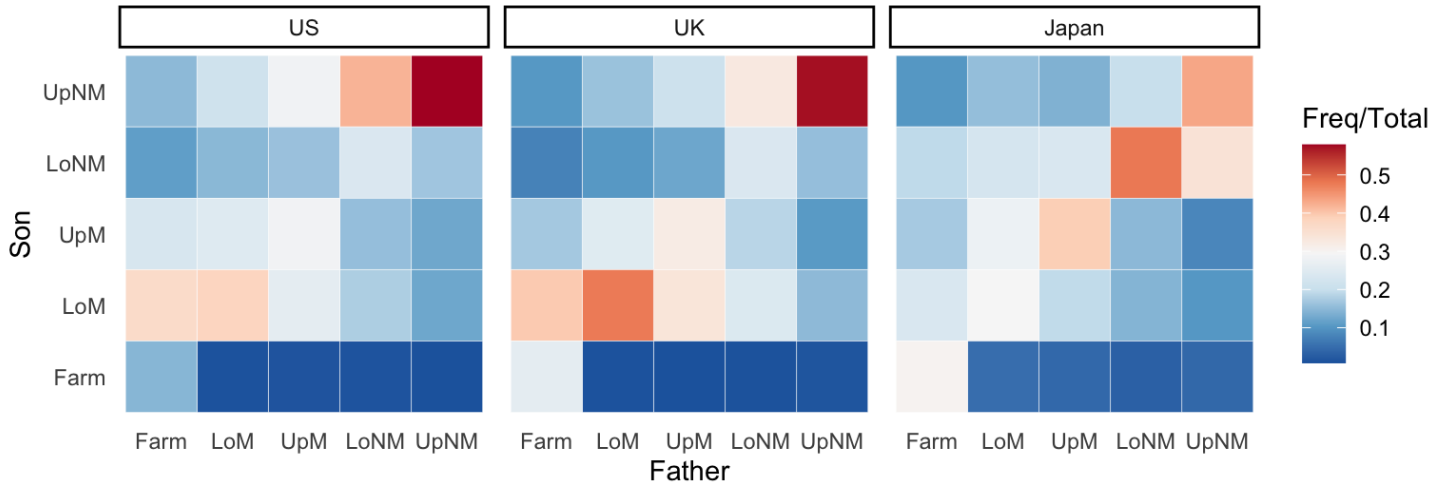


```
g + scale_fill_viridis_c(end = .9)
```



# RColorBrewer

```
g + scale_fill_distiller(palette = "RdBu")
```



```
ggplot(mydata3, aes(x = Father, y = Son)) +
  geom_tile(aes(fill = (Freq/Total)), color = "white") +
  coord_fixed() +
  scale_fill_gradient2(low = "black", mid = "white",
                       high = "red", midpoint = .2) +
  facet_wrap(~Country) + theme_heat
```