Homework 6

STAT 5014

Adeline Guthrie

October 9, 2019

Problem 3

```
# Part a
calc_prop <- function (vect){
  return(sum(vect)/length(vect))
}

# Part b
set.seed(12345)
P4b_data <- matrix(rbinom(10, 1, prob = (30:40)/100), nrow = 10, ncol = 10, byrow = FALSE)

# Part c
cprops <- apply(P4b_data, 2, calc_prop)
rprops <- apply(P4b_data, 1, calc_prop)
```

The proportion of successes for each row and column are exactly the same.

Column proportions: 0.6, 0.6, 0.6, 0.6, 0.6, 0.6, 0.6, 0.6, 0.6, 0.6

Row proportions: 1, 1, 1, 1, 0, 0, 0, 0, 1, 1

It appears the simulation created the same 10 coin flips for each row despite adjusting the "fairness" each time.

```
# Part d
calc_prop2 <- function (prb) {
  return(rbinom(10, 1, prob = prb))
}

probabilities <- (30:40)/100

new_matrix <- matrix(sapply(probabilities, calc_prop2), nrow = 10, ncol = 10, byrow = FALSE)

new_cprops <- apply(new_matrix, 2, calc_prop)
new_rprops <- apply(new_matrix, 1, calc_prop)
```

New matrix of coin flips:

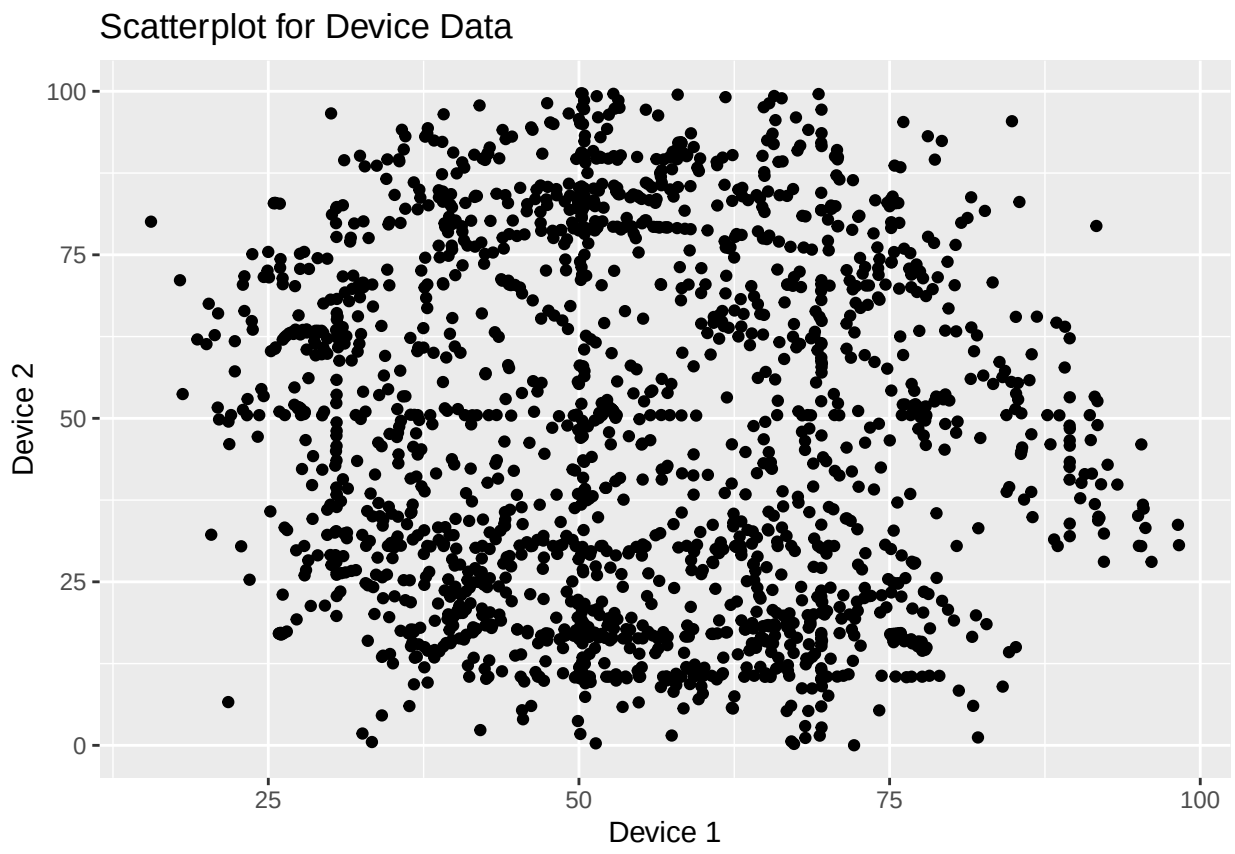
0	0	1	1	1	1	1	1	1	0
0	0	0	0	1	0	0	0	1	1
1	1	0	1	0	1	0	0	0	1
0	1	1	1	0	1	0	0	0	1
0	0	0	0	1	1	0	0	1	0
0	0	0	0	0	0	0	0	0	1

0	1	1	0	1	1	1	1	1	1
0	0	1	0	0	0	1	0	1	1
0	0	0	0	0	0	0	1	0	0
1	0	0	0	0	1	0	0	0	0

New column proportions: 0.2, 0.3, 0.4, 0.3, 0.4, 0.6, 0.3, 0.3, 0.5, 0.6

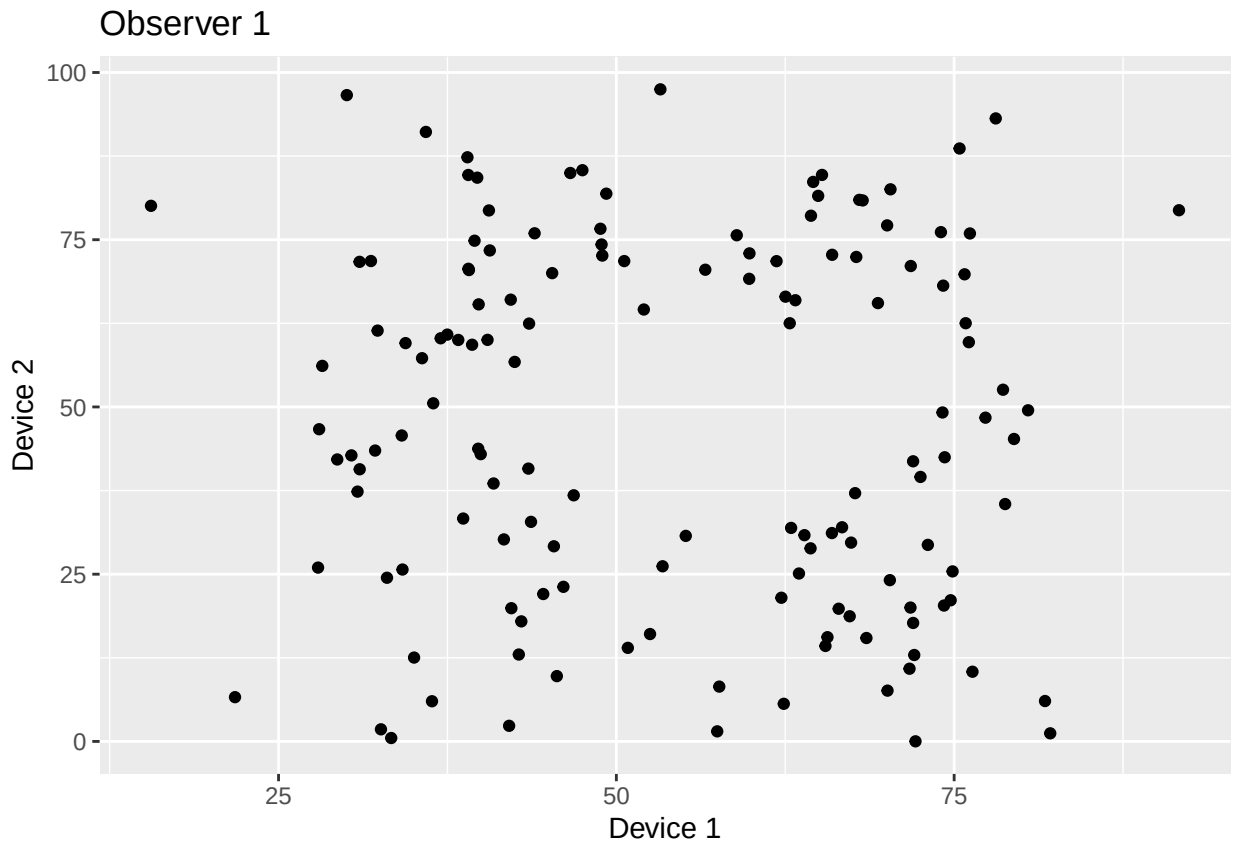
New row proportions: 0.7, 0.3, 0.5, 0.5, 0.3, 0.1, 0.8, 0.4, 0.1, 0.2

```
#Read in data
dat <- readRDS("/Users/Adeline/Documents/STAT_5014/HW4_data.rds")

#Tidy dataset
obsTidy <- dat %>%
  as_tibble() %>%
  arrange(Observer) %>%
  transmute(Observer = Observer, x = dev1, y = dev2)

create_scatterplot <- function(dat_frame, ttle, x_lab, y_lab) {
  ggplot(dat_frame, aes(x = x, y = y)) + geom_point() + labs(title = ttle, x = x_lab, y = y_lab)
}

# Create scatterplot for the entire dataset
create_scatterplot(obsTidy, "Scatterplot for Device Data", "Device 1", "Device 2")
```

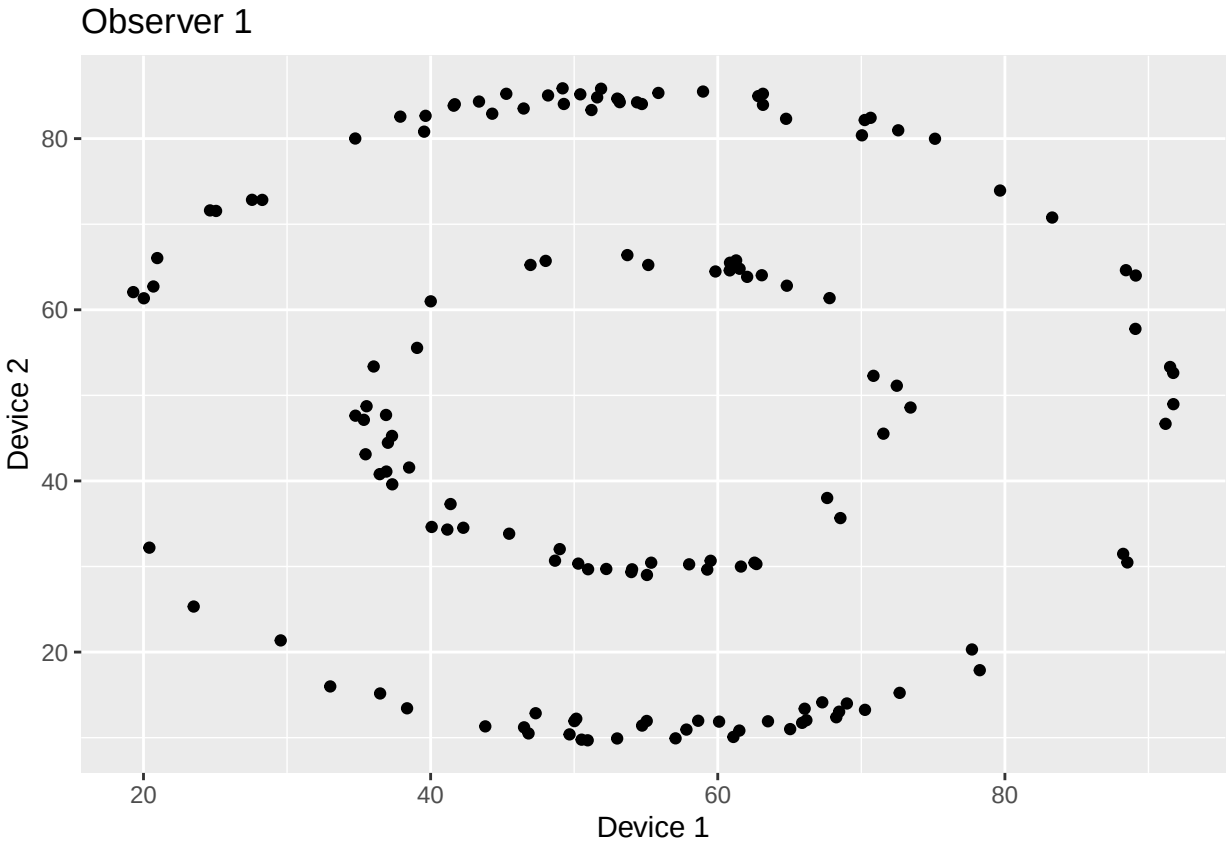


```r
# Create separate scatter plot for each observer
lapply(split(obsTidy, as.factor(obsTidy$Observer)), create_scatterplot, ttle = paste("Observer", obsTidy
```

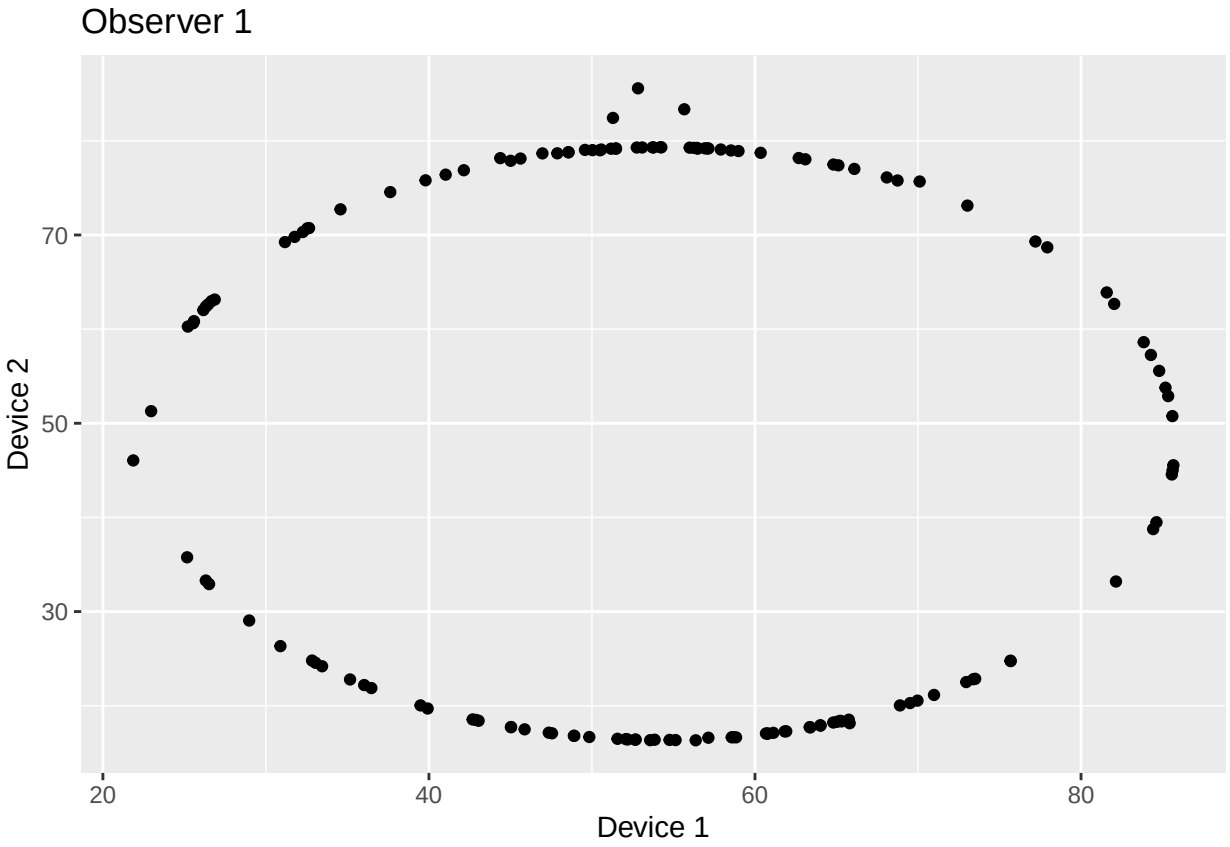
```
## $`1`
```



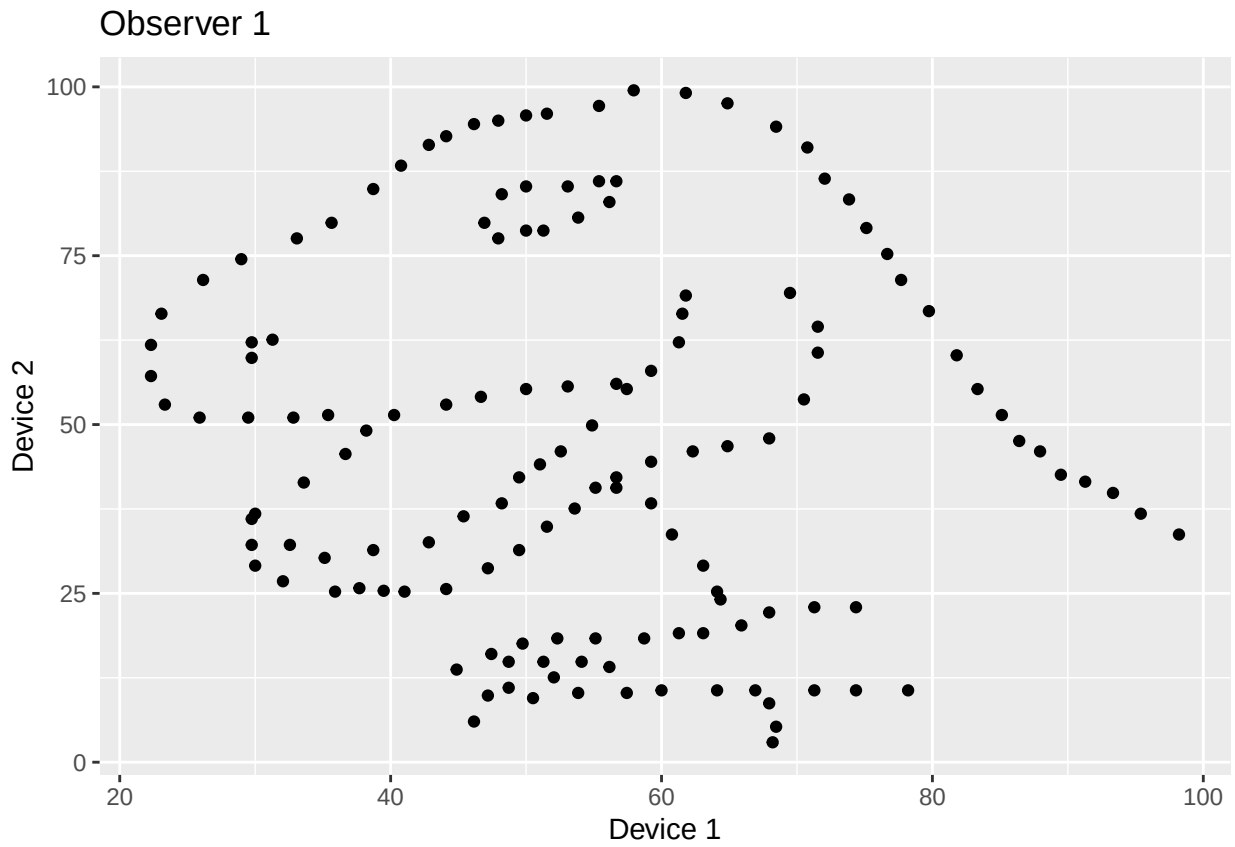

```
##
## $`2`
```



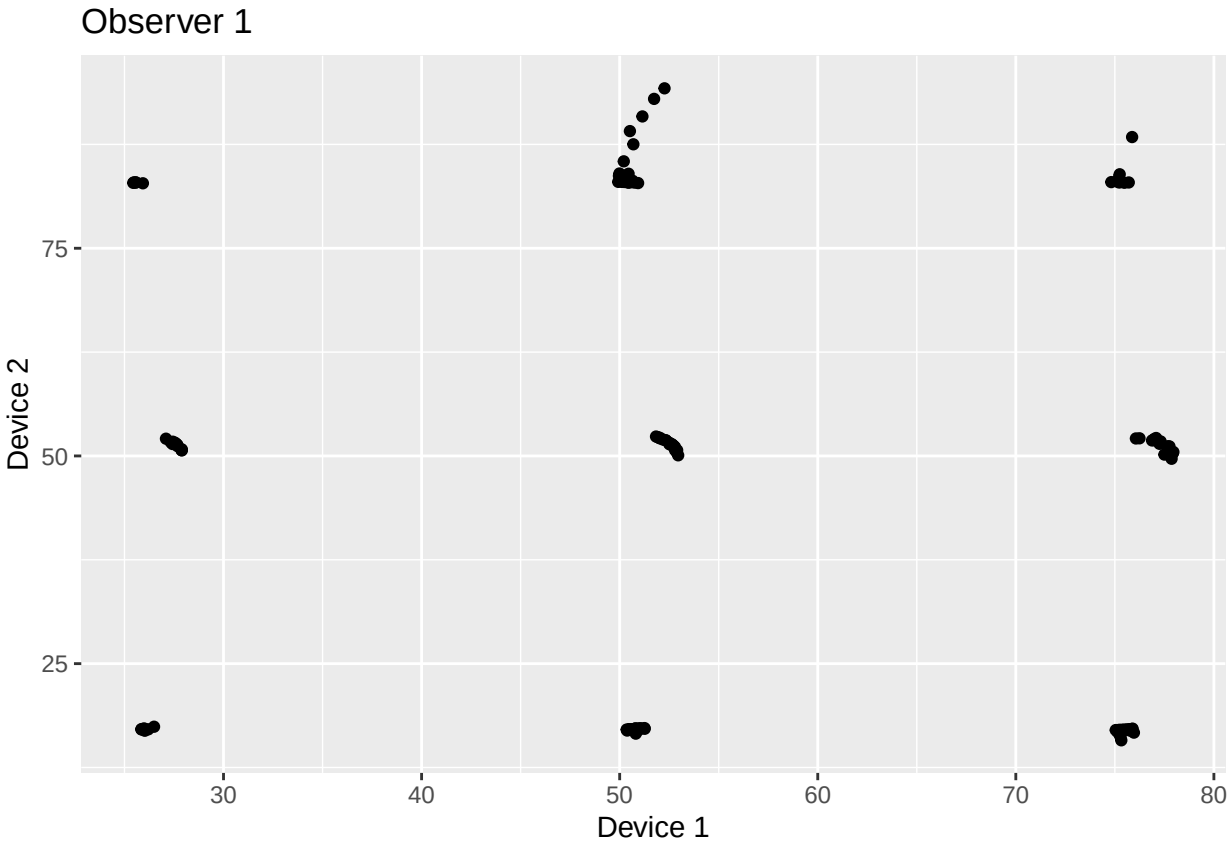

```
##
## $`3`
```



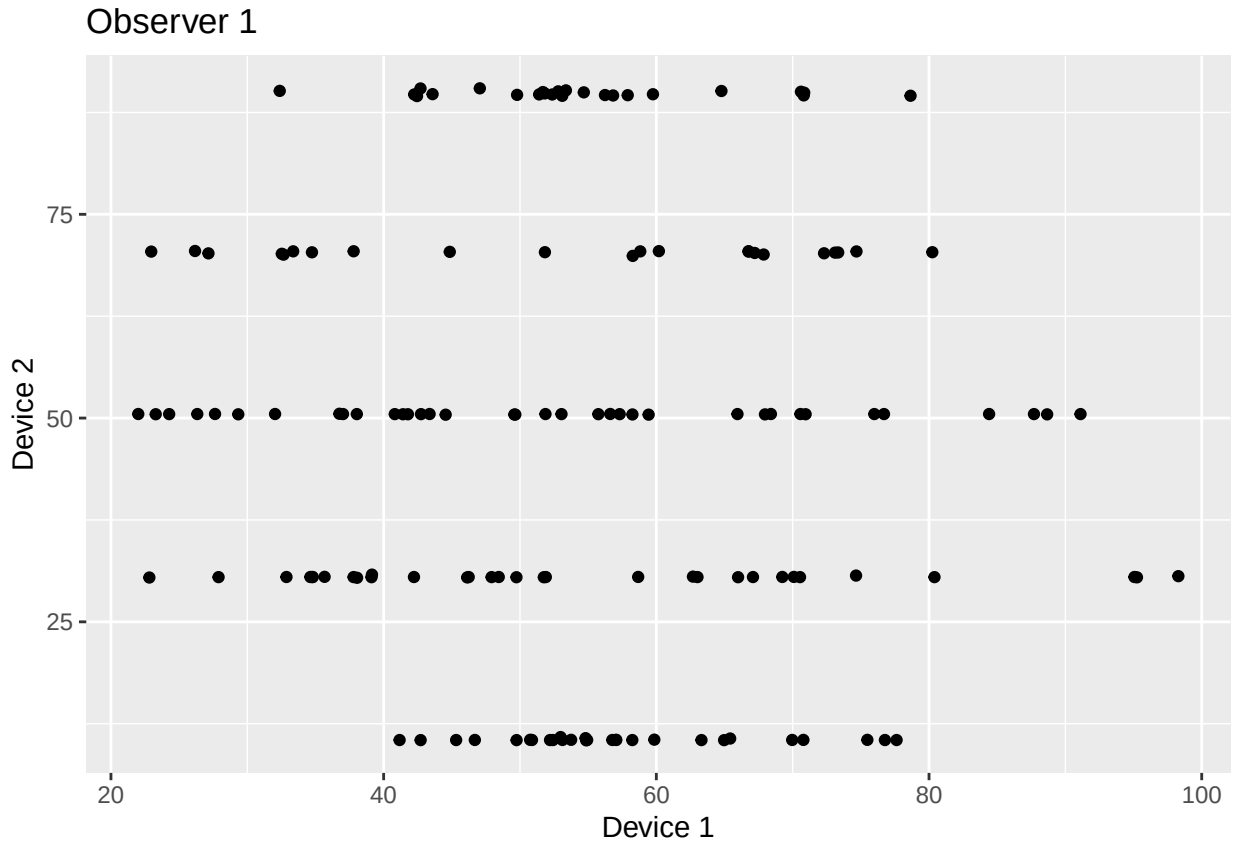

```
##
## $`4`
```



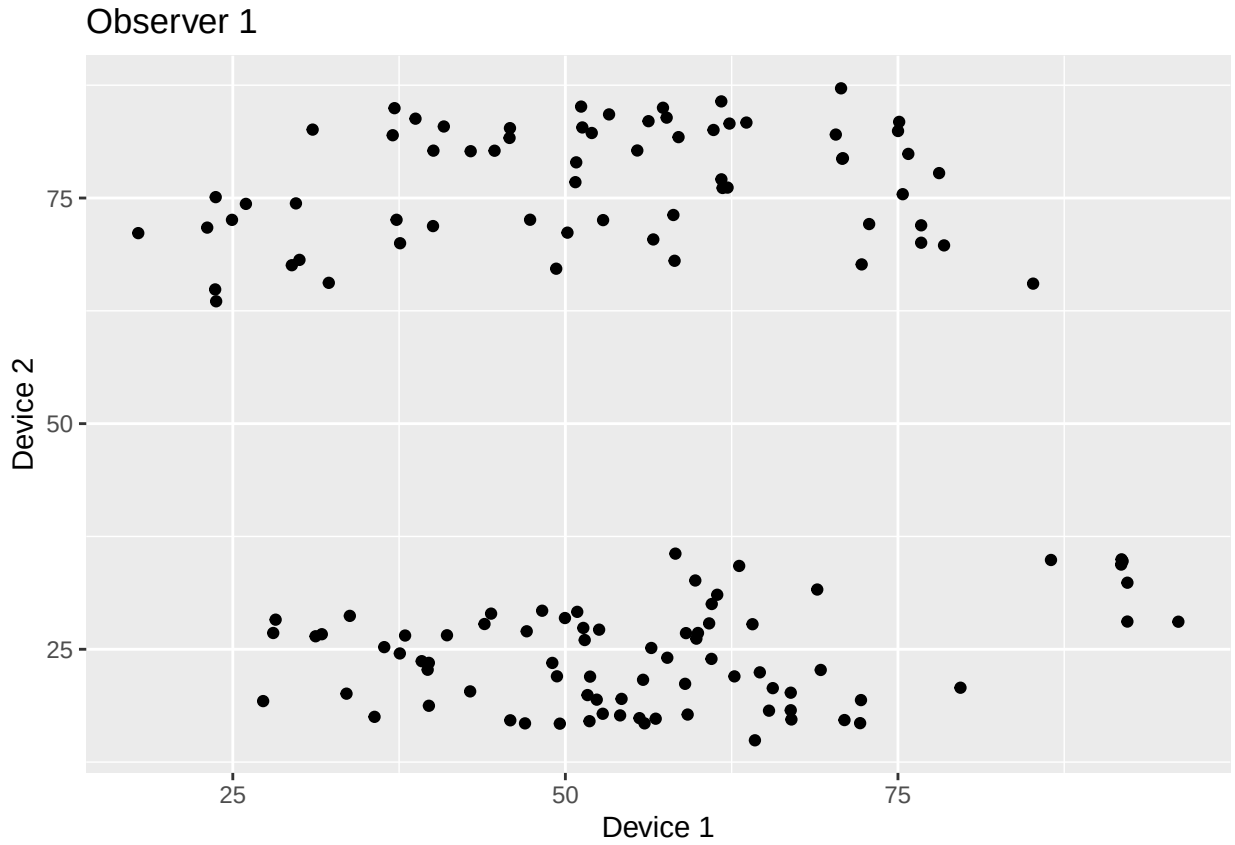

```
##
## $`5`
```



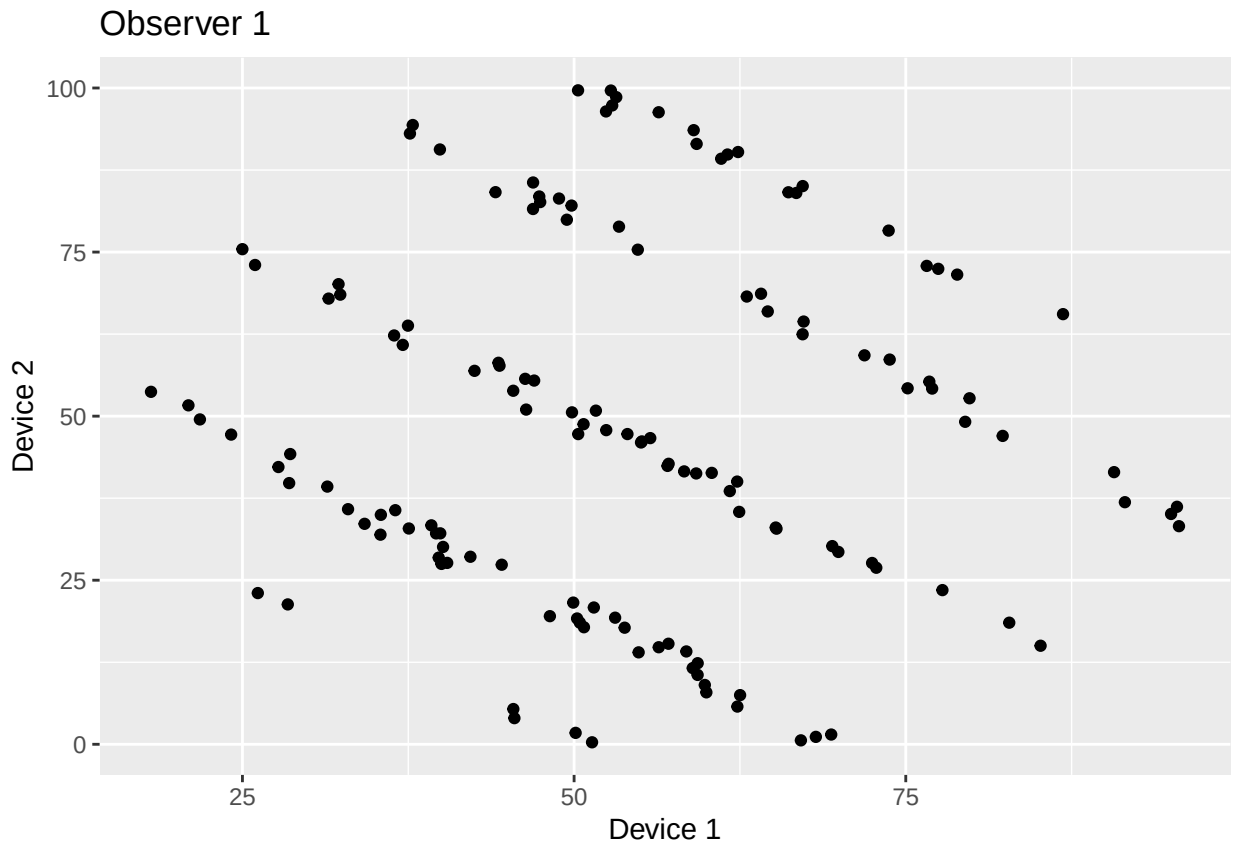

```
##
## $`6`
```



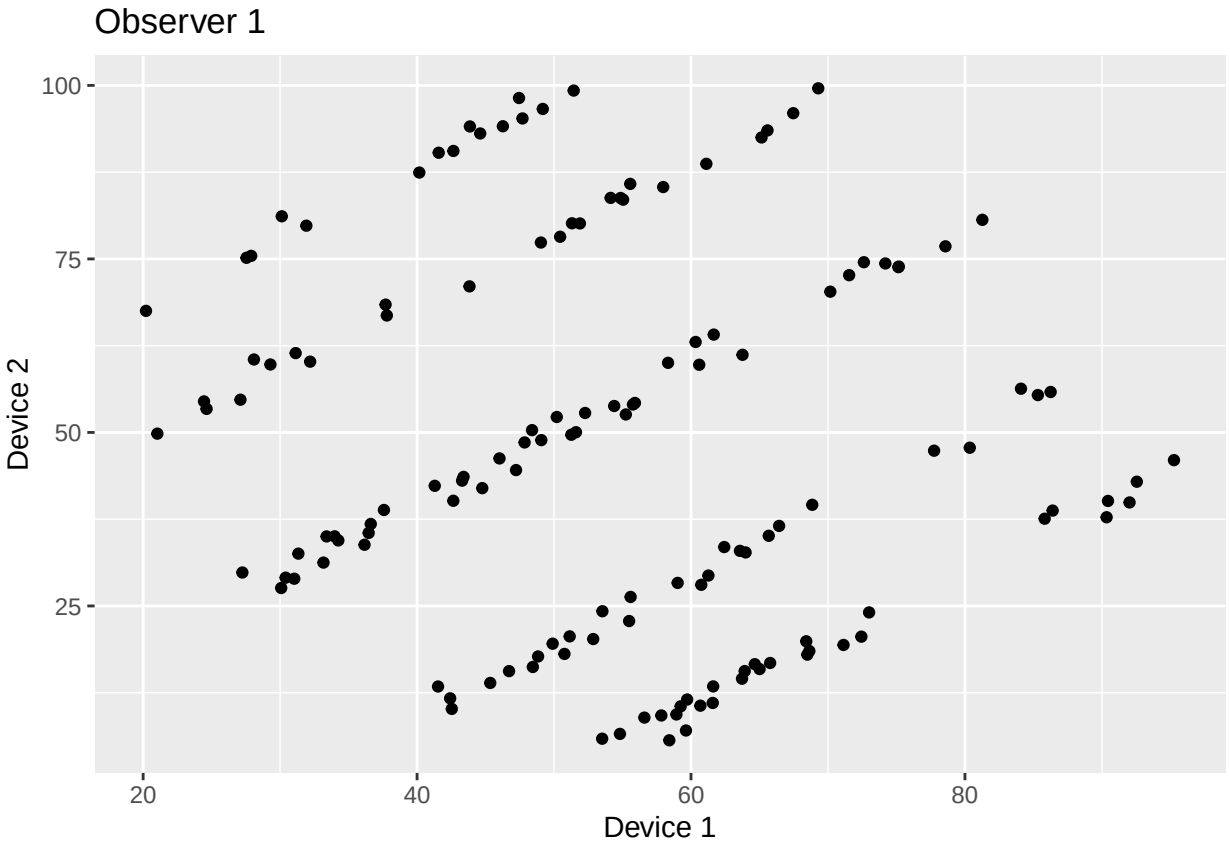

```
##
## $`7`
```



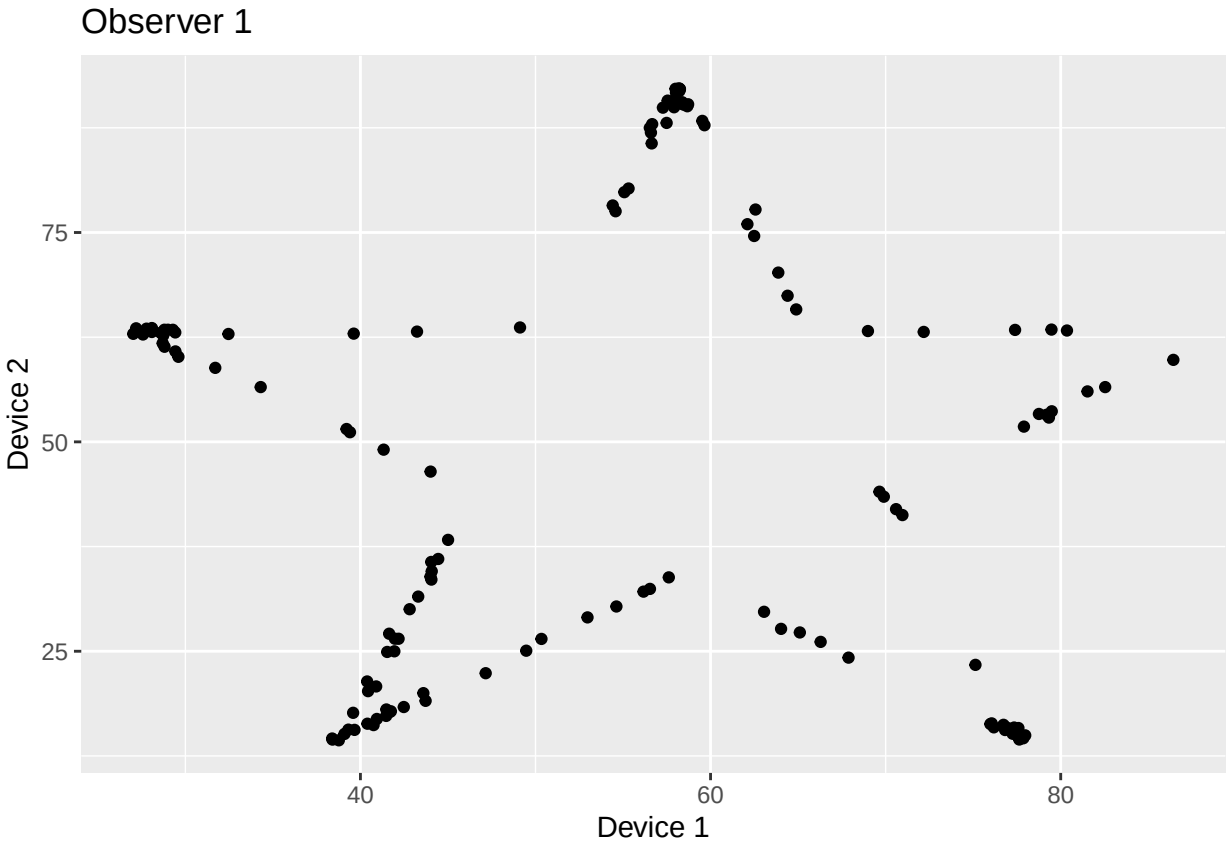

```
##
## $`8`
```



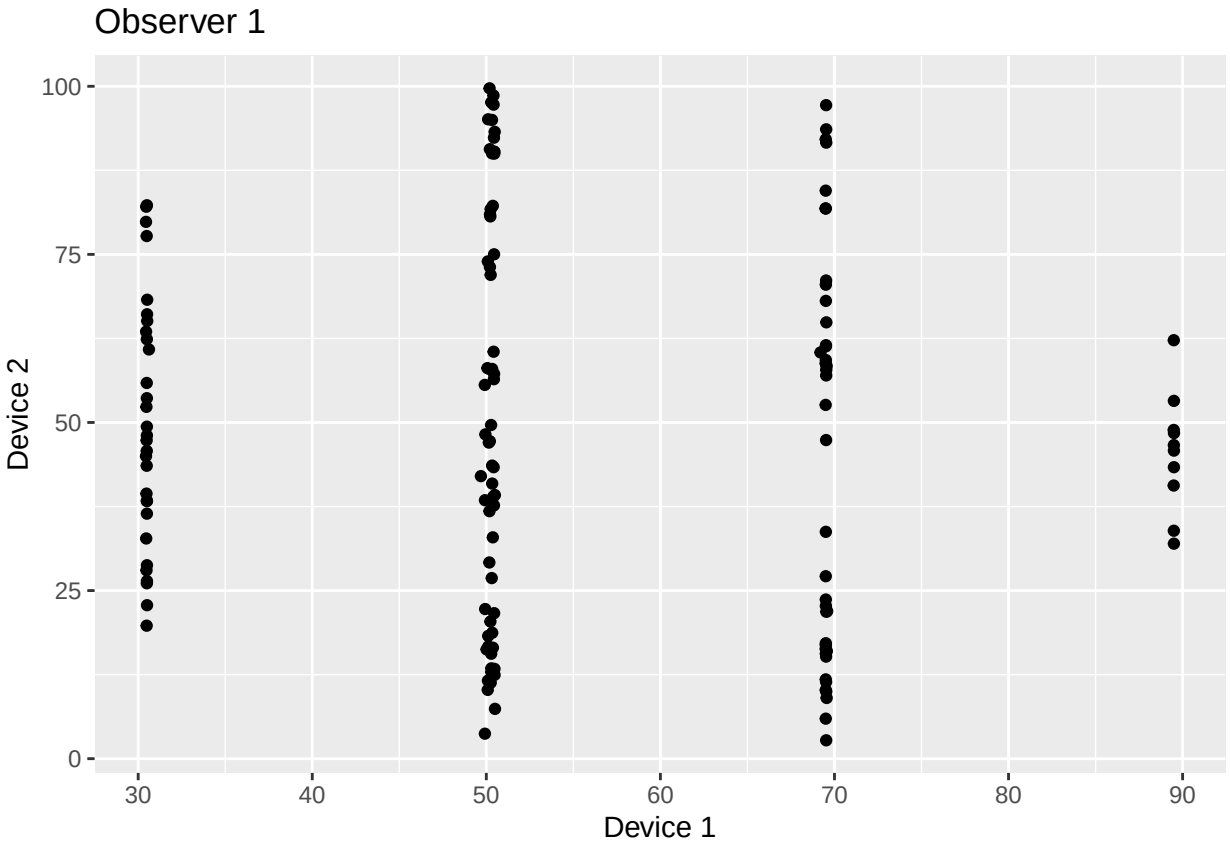

```
##
## $`9`
```



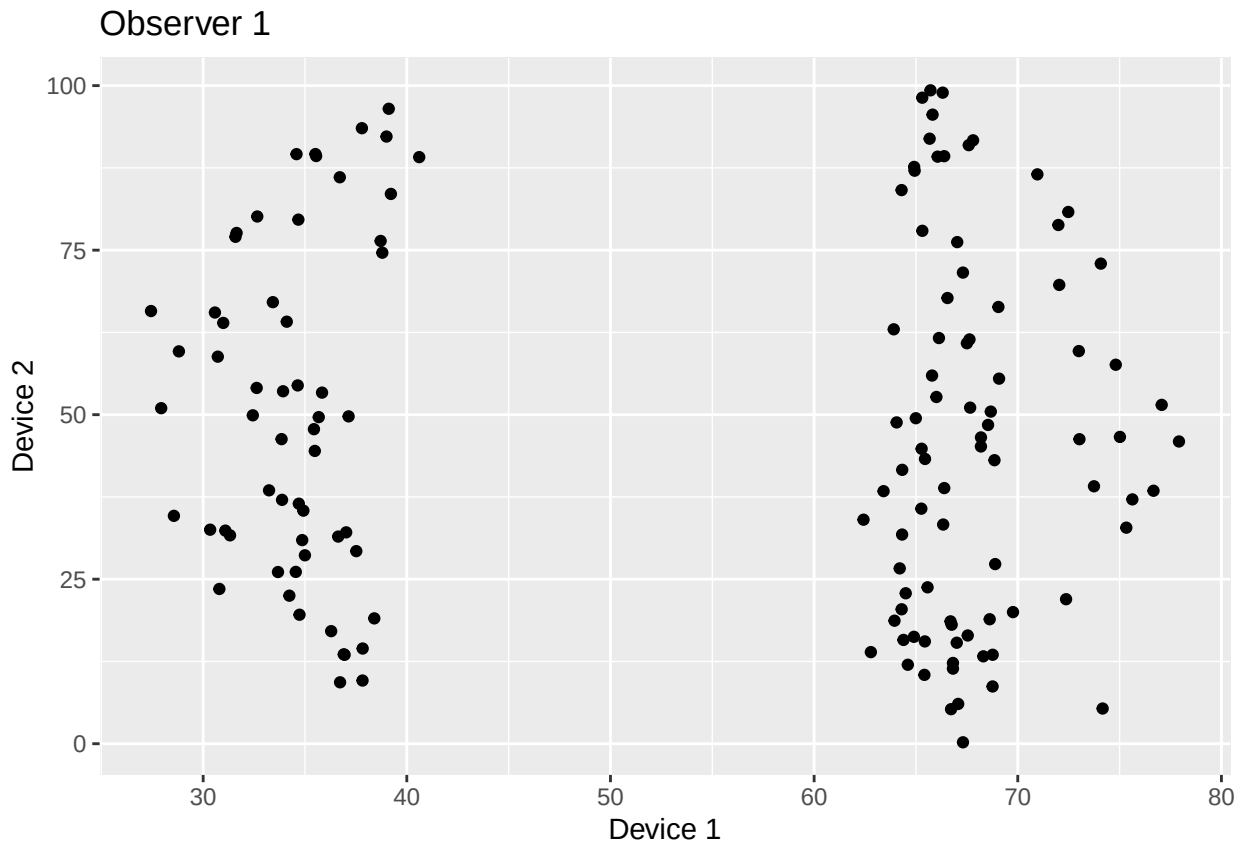

```
##
## $`10`
```



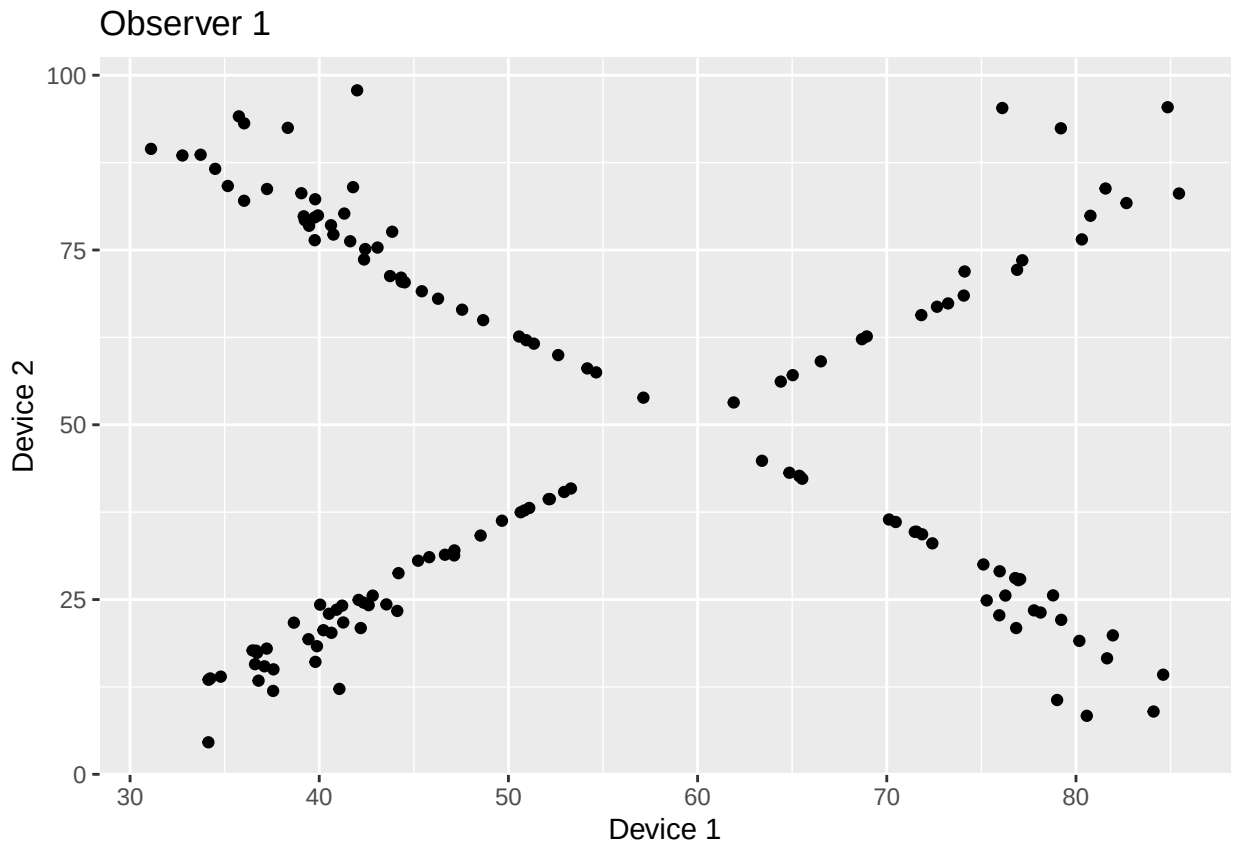

```
##
## $`11`
```




```
##
## $`12`
```




```
##
## $`13`
```

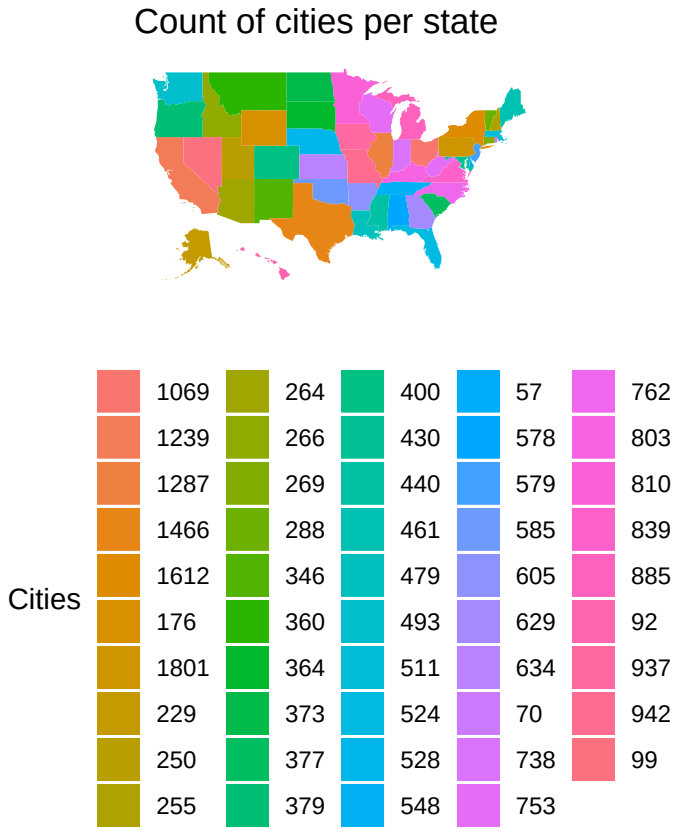



```
download("http://www.farinspace.com/wp-content/uploads/us_cities_and_states.zip",dest="us_cities_sta
unzip("us_cities_states.zip", exdir="C:/Users/Adeline/Documents/STAT_5014")

# Read in data
states <- fread(input = "./us_cities_and_states/states.sql",skip = 23,sep = "'",
                sep2 = ",", header = F, select = c(2,4))
cities <- fread(input = "./us_cities_and_states/cities.sql", skip = 25, sep ="'",
                sep2 = ",", header = FALSE, select = c(2, 4))#, 6, 8, 10, 12))

# Remove DC and PR
cities <- filter(cities, V4 != 'DC' & V4 != 'PR')
states <- filter(states, V4 != 'DC')

#Part b
# Create summary table
counts <- matrix(c(tolower(states$V2), rep(NA, 50)), nrow = 50, ncol = 2,
                 dimnames = list(c(), c("States", "Cities")))
ind <- 1
for (i in states$V4) {
  counts[ind,2] <- sum(cities$V4 == i)
  ind <- ind + 1
}

kable(counts)
```

States	Cities
alaska	229
alabama	578
arkansas	605
arizona	264
california	1239
colorado	400
connecticut	269
delaware	57
florida	524
georgia	629
hawaii	92
iowa	937
idaho	266
illinois	1287
indiana	738
kansas	634
kentucky	803
louisiana	479
massachusetts	511
maryland	430
maine	461
michigan	885
minnesota	810
missouri	942
mississippi	440
montana	360
north carolina	762
north dakota	373
nebraska	528
new hampshire	255
new jersey	579
new mexico	346
nevada	99
new york	1612
ohio	1069
oklahoma	585
oregon	379
pennsylvania	1801
rhode island	70
south carolina	377
south dakota	364
tennessee	548
texas	1466
utah	250
virginia	839
vermont	288
washington	493
wisconsin	753
west virginia	753
wyoming	176

```r
# Part c
letter_count <- data.frame(matrix(nrow = 50, ncol = 26))

getCount <- function(name, letter){
  temp <- unlist(strsplit(name, ""))
  count <- sum(sapply(tolower(temp), identical, y = letter))
  return(count)
}

for(i in 1:50){
  letter_count[i,] <- sapply(letters[1:26], getCount, name = states[i, 1])
}
```

```r
#https://cran.r-project.org/web/packages/fiftystater/vignettes/fiftystater.html


data("fifty_states") # this line is optional due to lazy data loading
crimes <- data.frame(state = tolower(rownames(USArrests)), USArrests)
counts <- data.frame(counts)

# First Map
# map_id creates the aesthetic mapping to the state name column in your data
p <- ggplot(counts, aes(map_id = States)) +
  # map points to the fifty_states shape data
  geom_map(aes(fill = Cities), map = fifty_states) +
  expand_limits(x = fifty_states$long, y = fifty_states$lat) +
  coord_map() +
  scale_x_continuous(breaks = NULL) +
  scale_y_continuous(breaks = NULL) +
  labs(title = "Count of cities per state", x = "", y = "") +
  theme(legend.position = "bottom",
        panel.background = element_blank())

p
```




```
# Filter out states that do not have more than 3 occurances of any letter in their name
letter_count2 <- letter_count %>%
  bind_cols("state" = tolower(states$V2)) %>%
  filter_at(vars(starts_with("X")),any_vars(. > 2))

# Second Map
p2 <- ggplot(letter_count2, aes(map_id = state)) +
  geom_map(aes(fill = state), map = fifty_states) +
  expand_limits(x = fifty_states$long, y = fifty_states$lat) +
  coord_map() +
  scale_x_continuous(breaks = NULL) +
  scale_y_continuous(breaks = NULL) +
  labs(title = "States that have more than 3 occurances of any letter in their name",
       x = "", y = "") +
  theme(legend.position = "bottom",
        panel.background = element_blank())

p2
```

States that have more than 3 occurances of any letter in their name