Effective Data Visualization

Neda Sadeghi, Ph.D.
Data Scientist, Section on Clinical and Computation Psychiatry
National Institute of Mental Health, National Institutes of Health
April 5, 2021

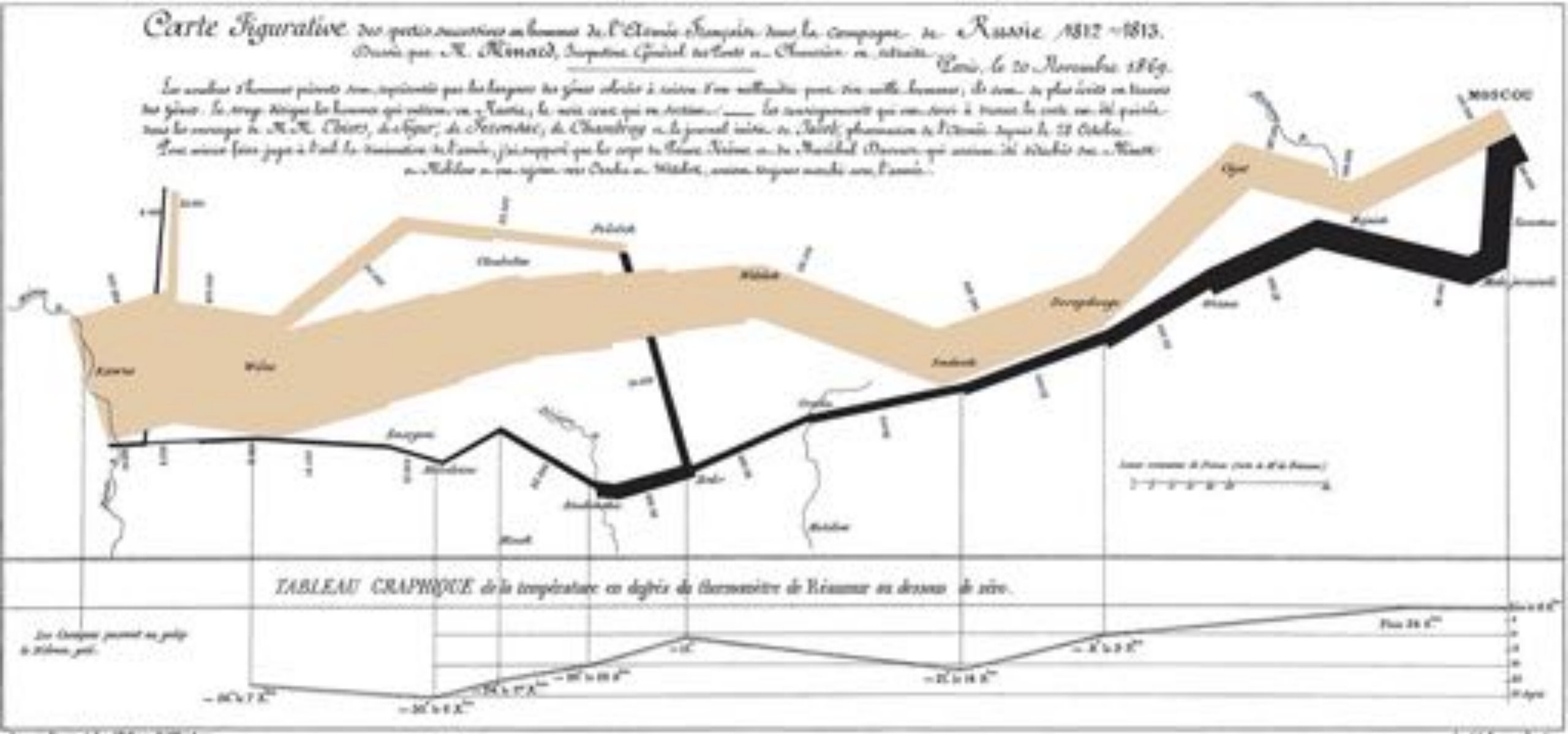
Overview

- Visualization principles
- R – ggplot2
- COVID effects on healthy and depressed adolescents
- Visualization of brain imaging data

Why data visualization?

- Communication
- Exploratory data analysis – scientific discovery process
- Can help detect bias, errors, unexpected results

Minard's Visualization of Napolean's 1812 March



How many times the word data appears in this paragraph?

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

Source: https://www.tableau.com/learn/articles/data-visualization

How times the word data appears in this paragraph?

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

Source: https://www.tableau.com/learn/articles/data-visualization

Anscombe’s quartet

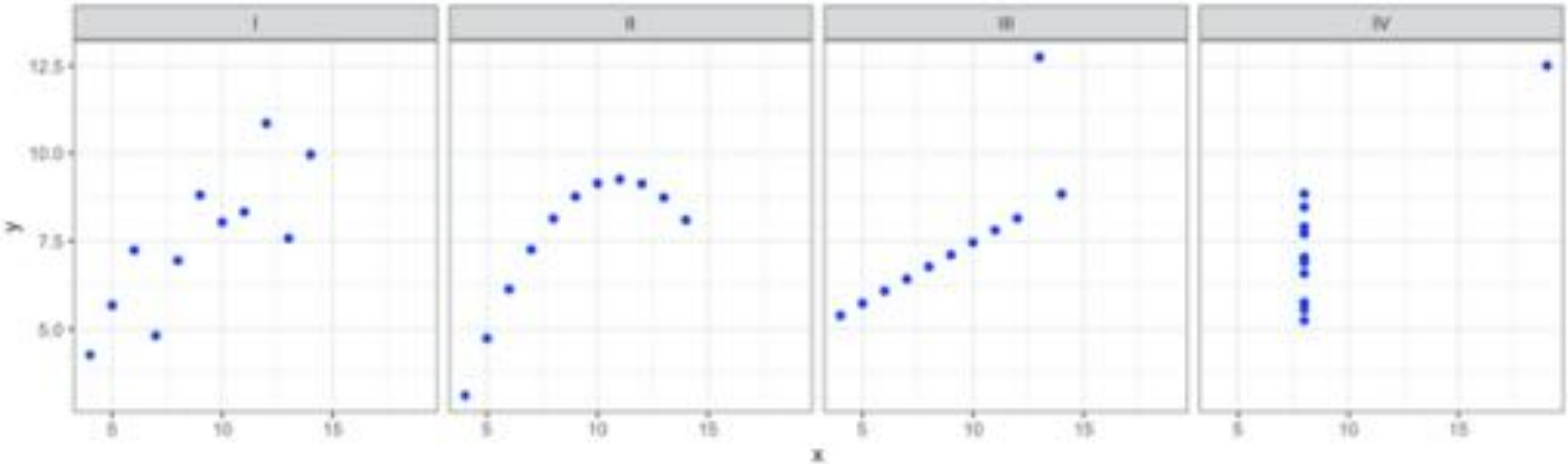
Dataset I		Dataset II		Dataset III		Dataset IV	
x	y	x	y	x	y	x	y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

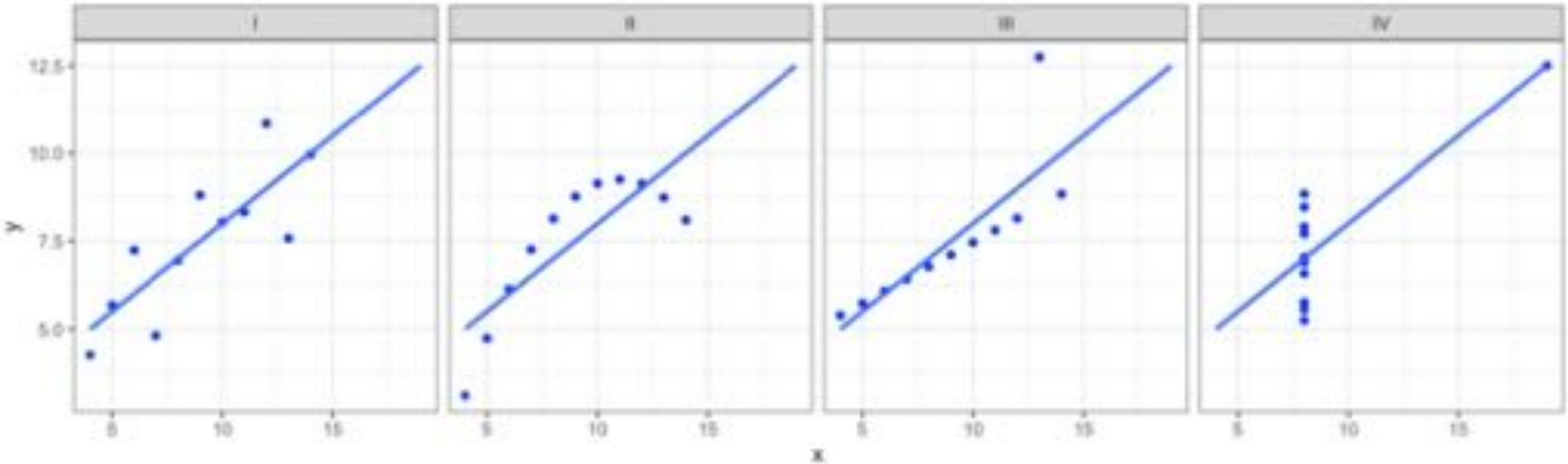
Data downloaded from: https://gist.github.com/ericbusboom
https://en.wikipedia.org/wiki/Anscombe%27s_quartet

Basic Statistics

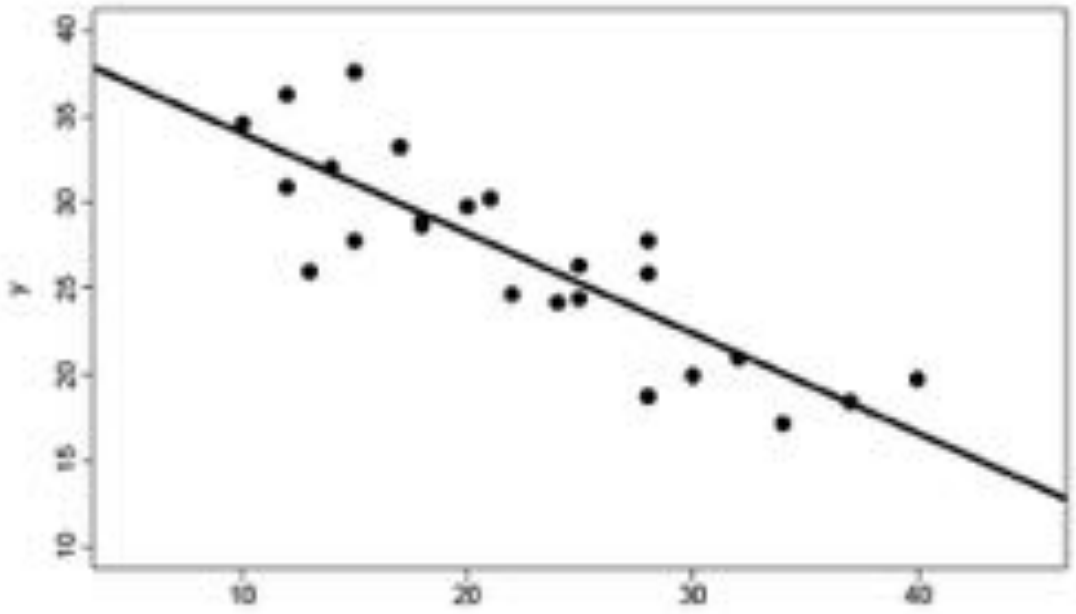
	Dataset I	Dataset II	Dataset III	Dataset IV
Mean of X	9	9	9	9
Variance of X	11	11	11	11
Mean of Y	9	9	9	9
Variance of Y	7.5	7.5	7.5	7.5
Correaltion	0.816	0.816	0.816	0.816

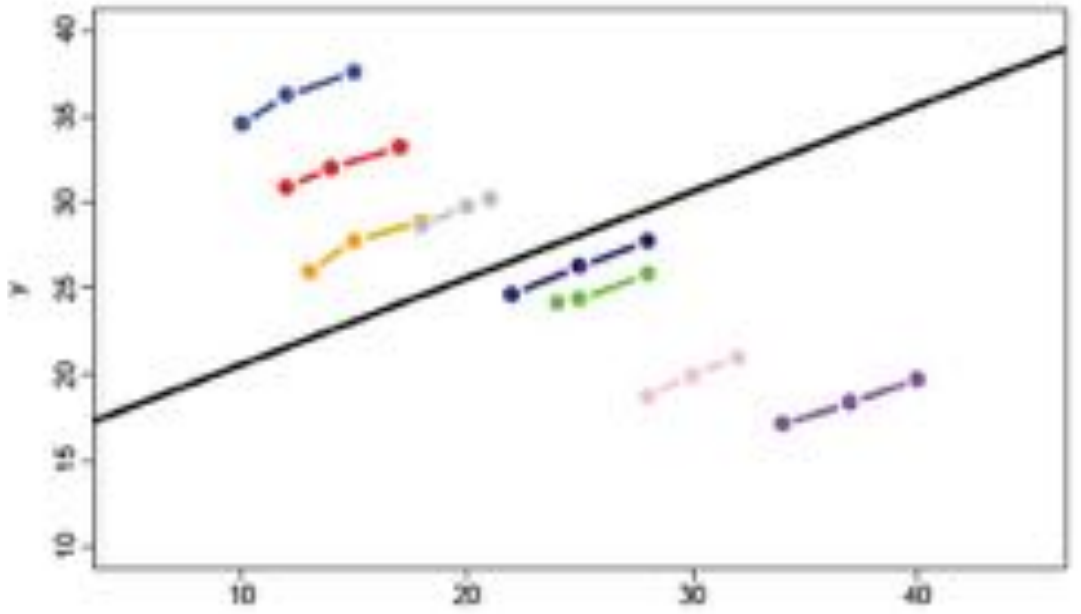
All four datasets have the same mean, variance, and correlation


I
II
III
IV
12.5
10.0
7.5
5.0
y
5
10
15
x


I
II
III
IV
12.5
10.0
7.5
5.0
y
5
10
15
x

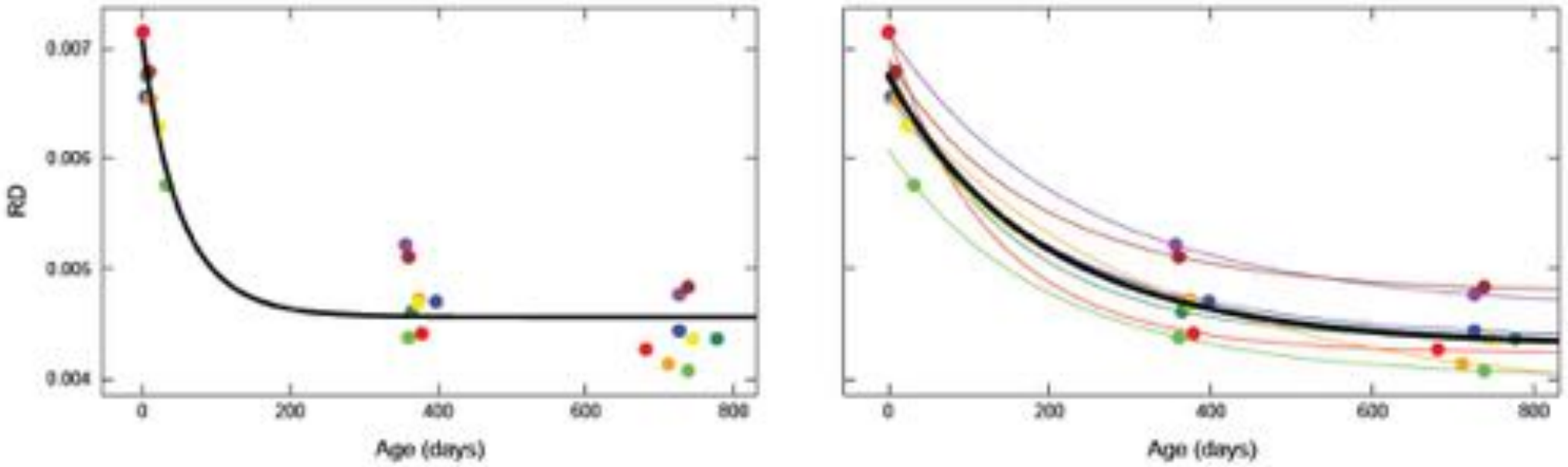
Increase or a decrease over time?





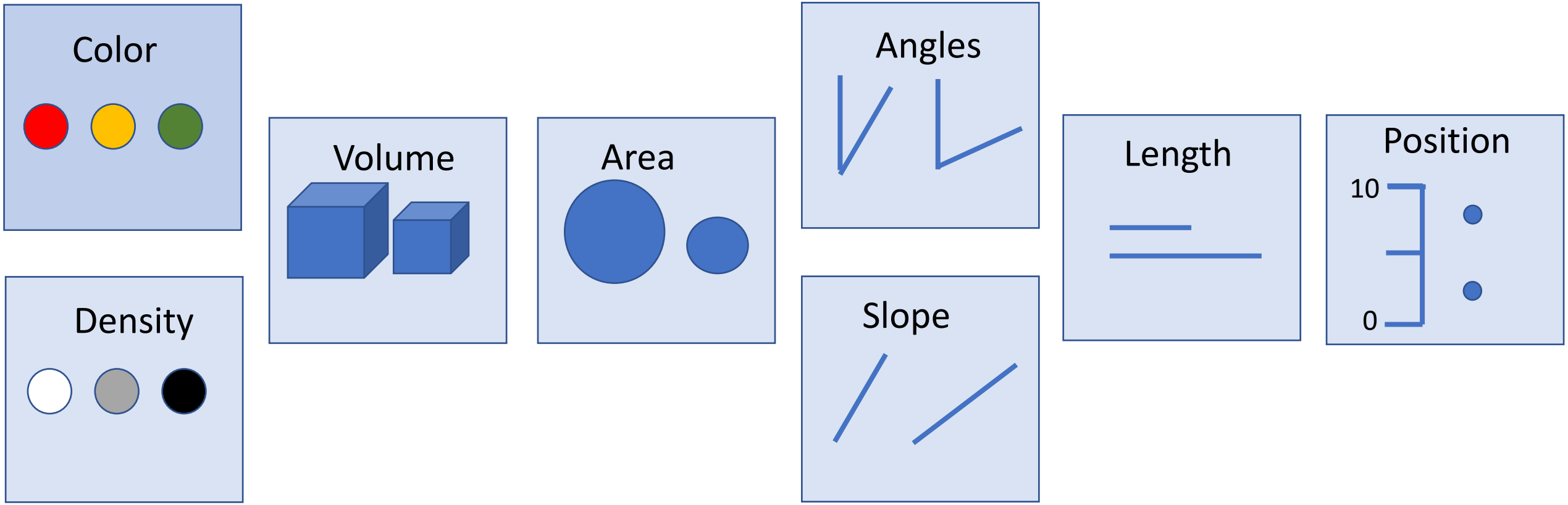
Sadeghi et al., NeuroImage, 2013

Modeling of Longitudinal MRI in Early Brain Development



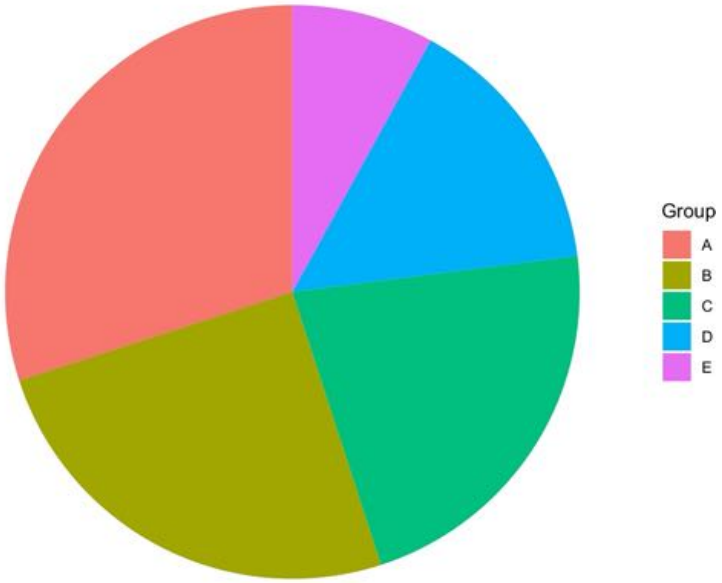
Sadeghi et al., NeuroImage, 2013

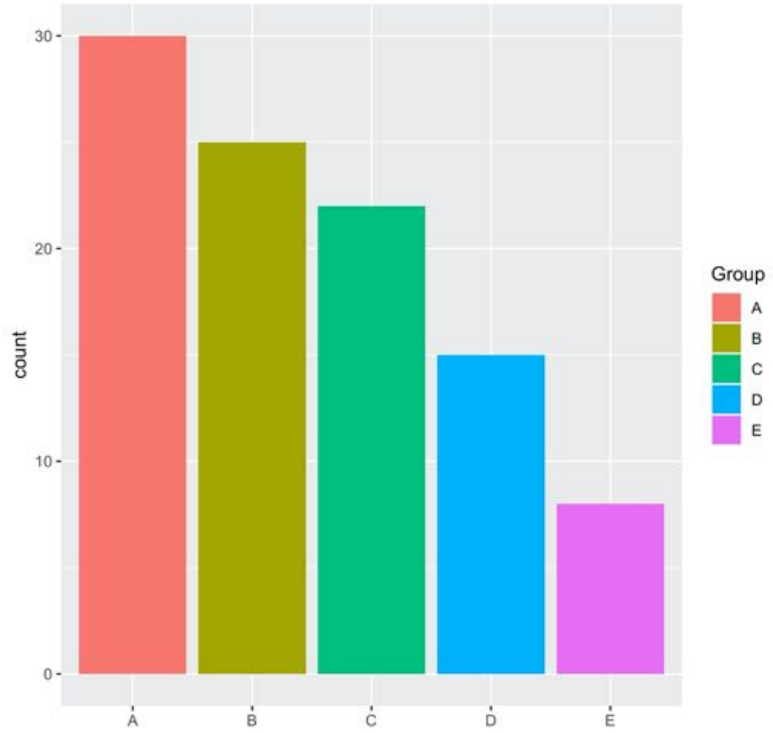
Foundational principles: visual encoding channels



Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. William S. Cleveland and Robert McGill. Inspired by Miriah Meyer's talk.

Pie chart or Bar plot?





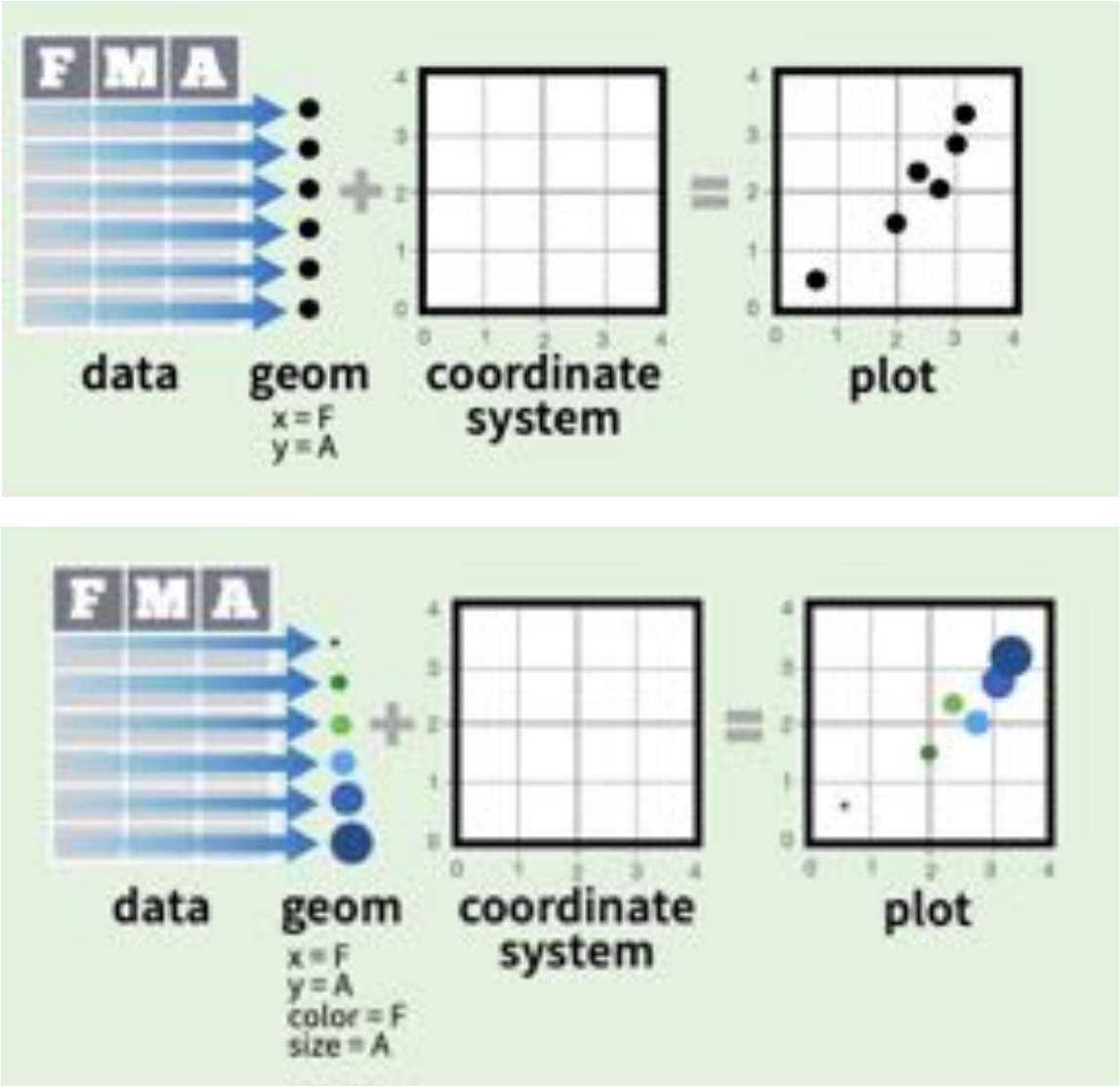
Why R and ggplot2?

- Excel
- R
- Python
- D3/interactive web apps

Why R and ggplot2?

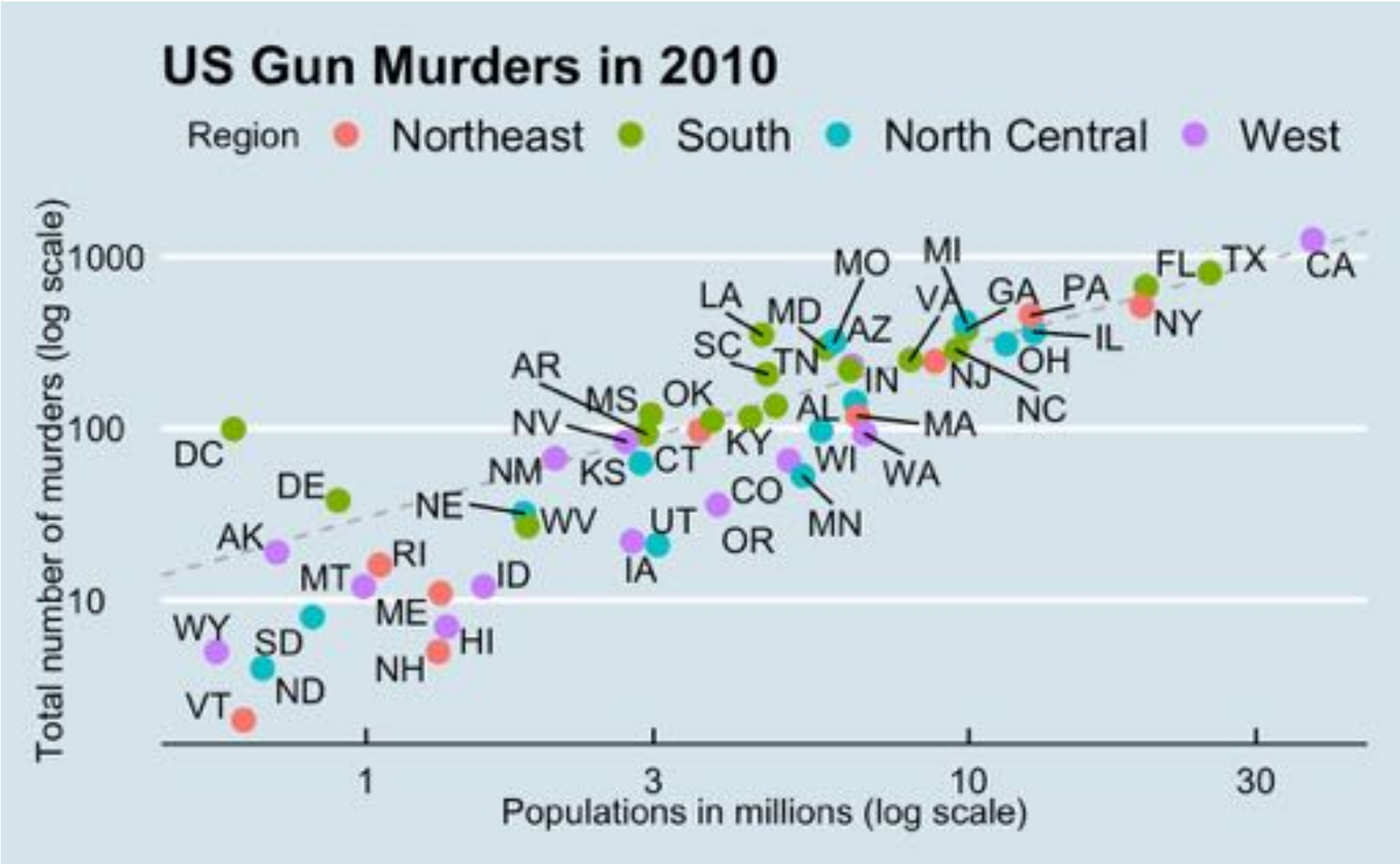
- Excel
- R - ggplot2
- Python
- D3/interactive web apps

ggplot2



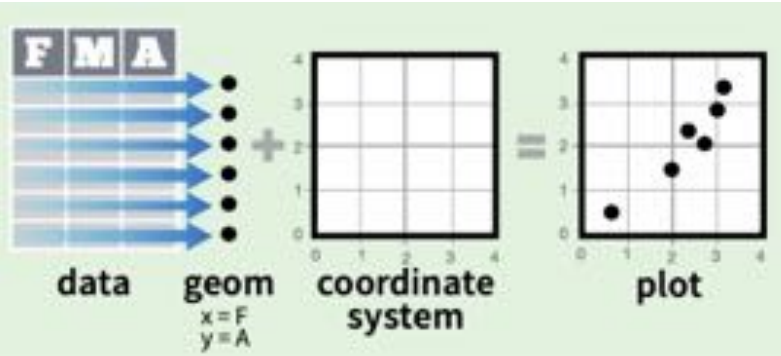
Data: rows are observations, columns are variables
geom: visual marks that represent data points

Source: rstudio.com



Source: https://rafalab.github.io/dsbook/ggplot2.html
Examples from Rafael A. Irizarry's website

```
install.packages("dslabs")
library(dslabs)
data(murders)
```




```
> head(murders)
```

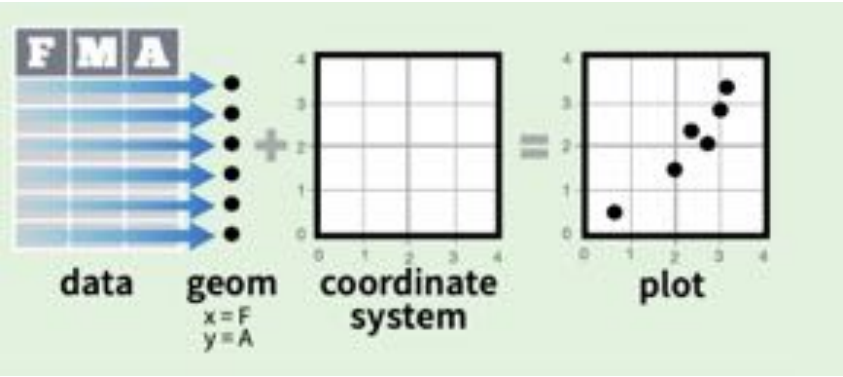
	state	abb	region	population	total
1	Alabama	AL	South	4779736	135
2	Alaska	AK	West	710231	19
3	Arizona	AZ	West	6392017	232
4	Arkansas	AR	South	2915918	93
5	California	CA	West	37253956	1257
6	Colorado	CO	West	5029196	65

Variables

Observations

Source: https://rafalab.github.io/dsbook/ggplot2.html
Examples from Rafael A. Irizarry's website

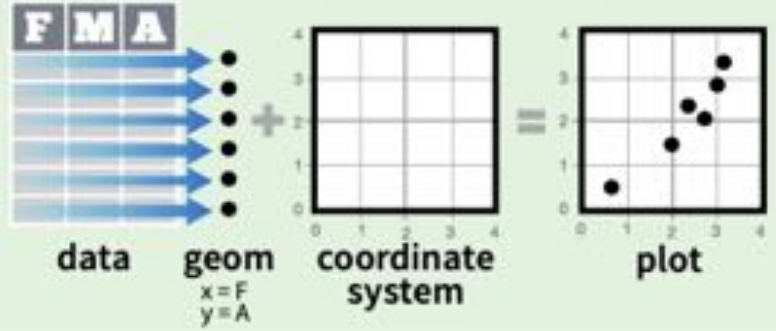
Geometries



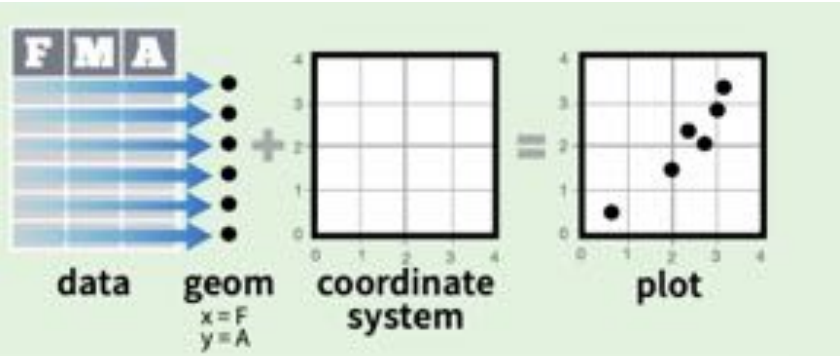
- ggplot2 works with layers
- Layers can define:
  - Geometry
  - Statistics
  - Labels
- To add layers use +

  ggplot(data) + layer1 + layer2 + …
- First layer usually is the geometry
  - geom_point (scatter plot)
  - geom_bar (bar plot)
  - geom_histogram

```
install.packages("dslabs")
library(dslabs)
data(murders)
```



```
install.packages("dslabs")
library(dslabs)
data(murders)
ggplot(data = murders)
```

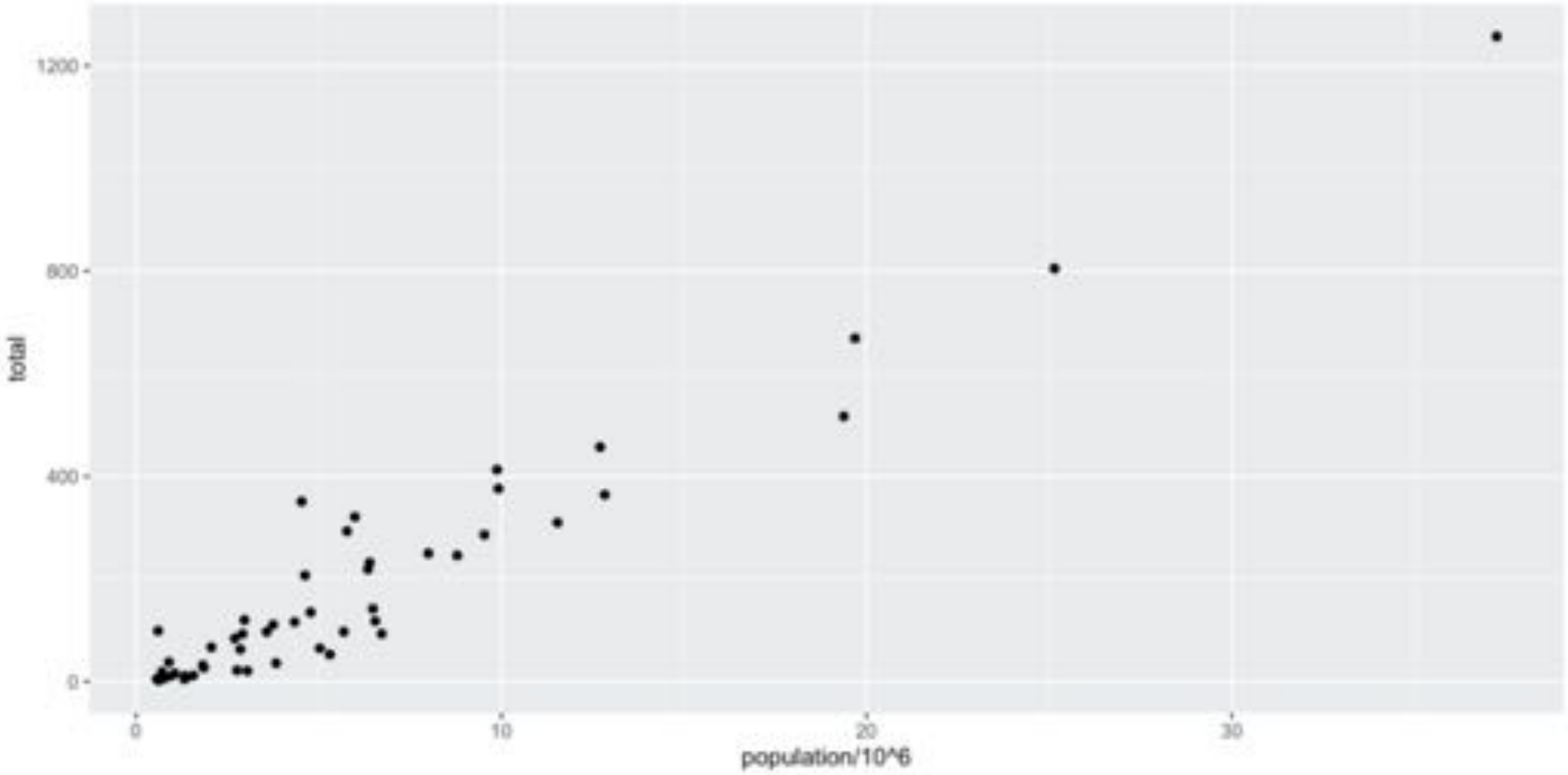


```
?geom_point
```

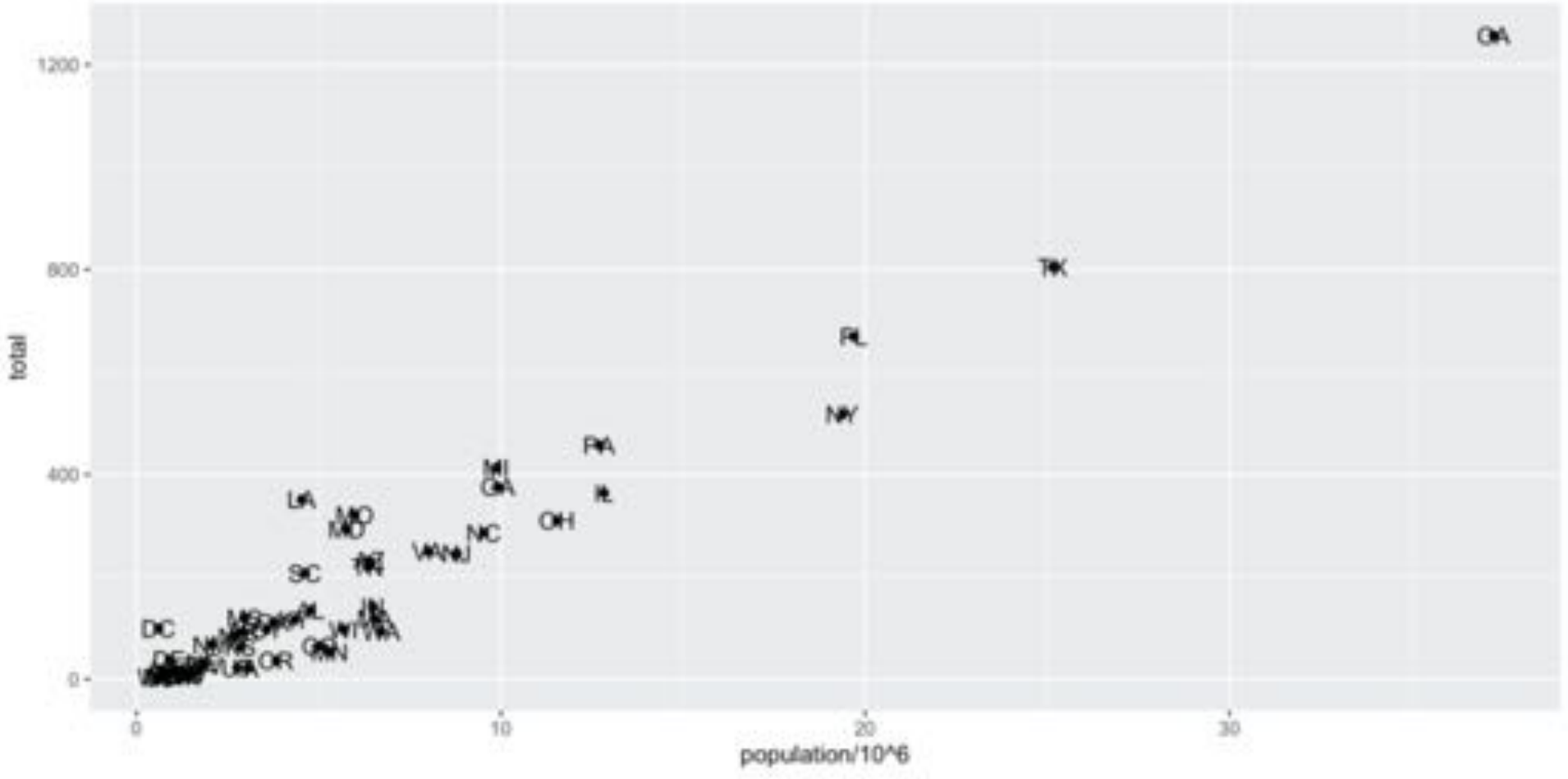
- geom_point() understands the following aesthetics (required aesthetics are in bold):
- **x**
- **y**
- alpha
- colour
- fill
- group
- shape
- size
- stroke

Aesthetic mappings describe how variables in the data are mapped to visual properties (**aesthetics**) of geoms.

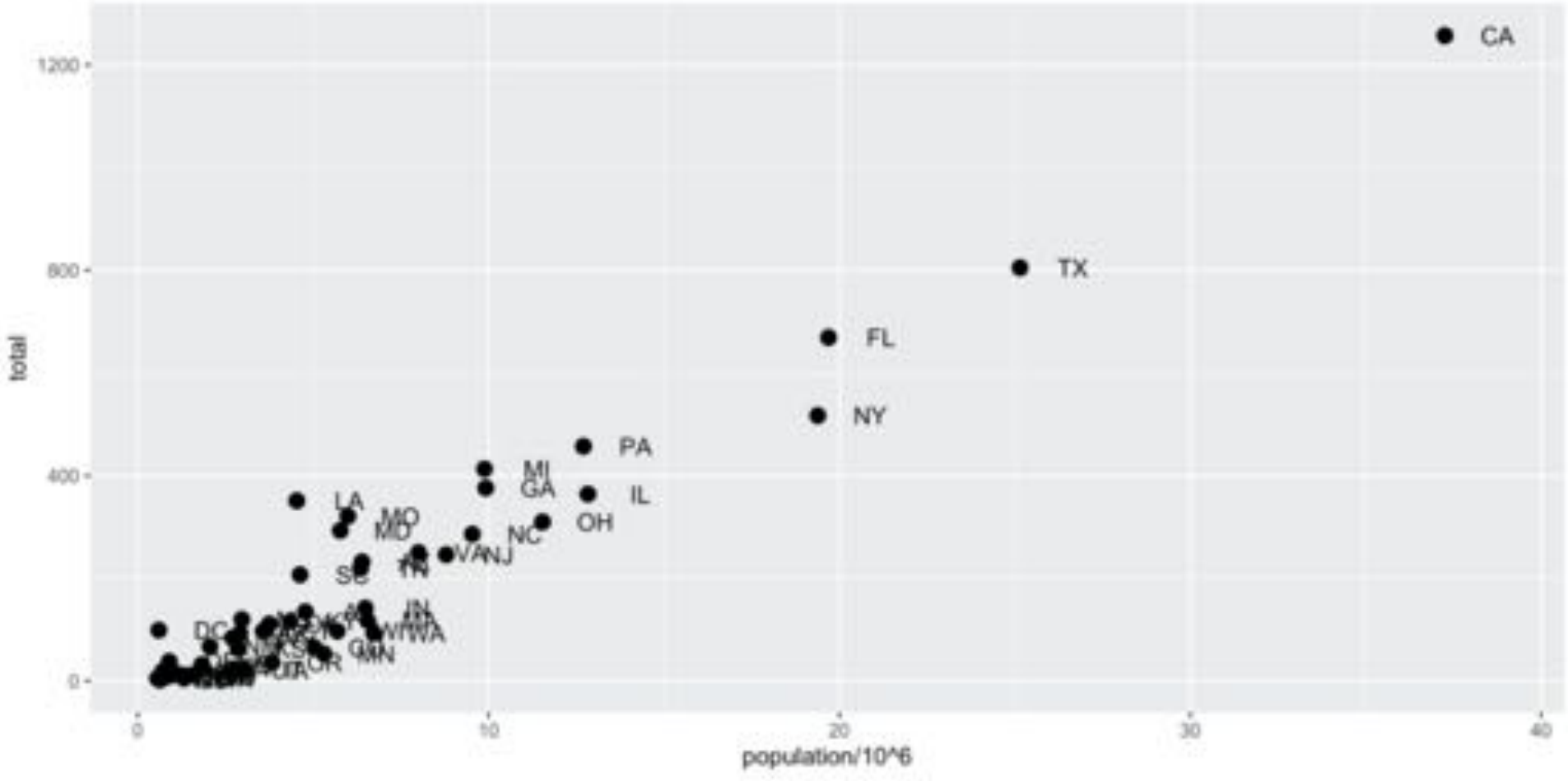
```
ggplot(data = murders) +
geom_point(aes(x = population/10^6, y = total))
```



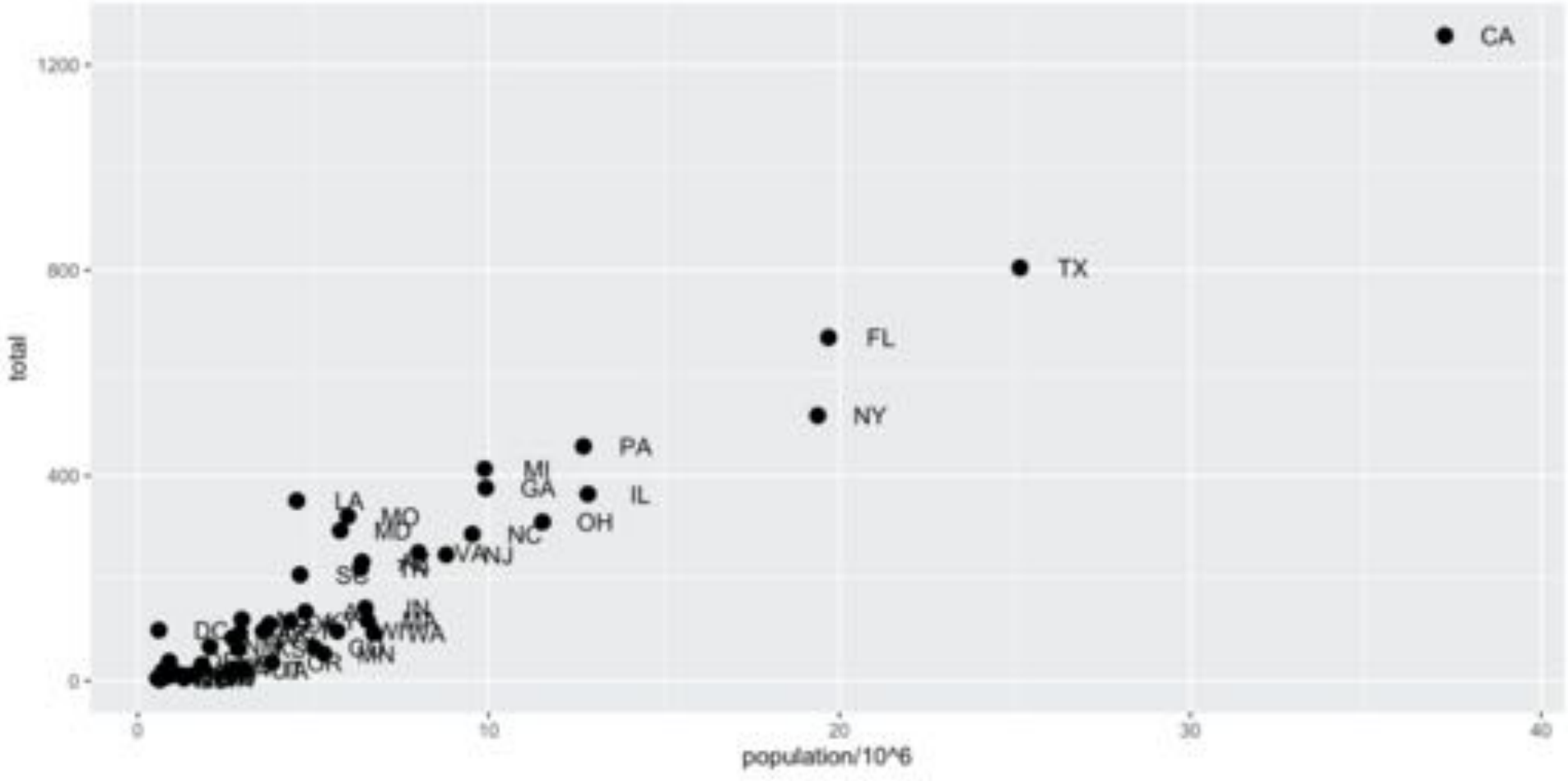
```
ggplot(data = murders) +
geom_point(aes(x = population/10^6, y = total))+
geom_text(aes(population/10^6, total, label = abb))
```



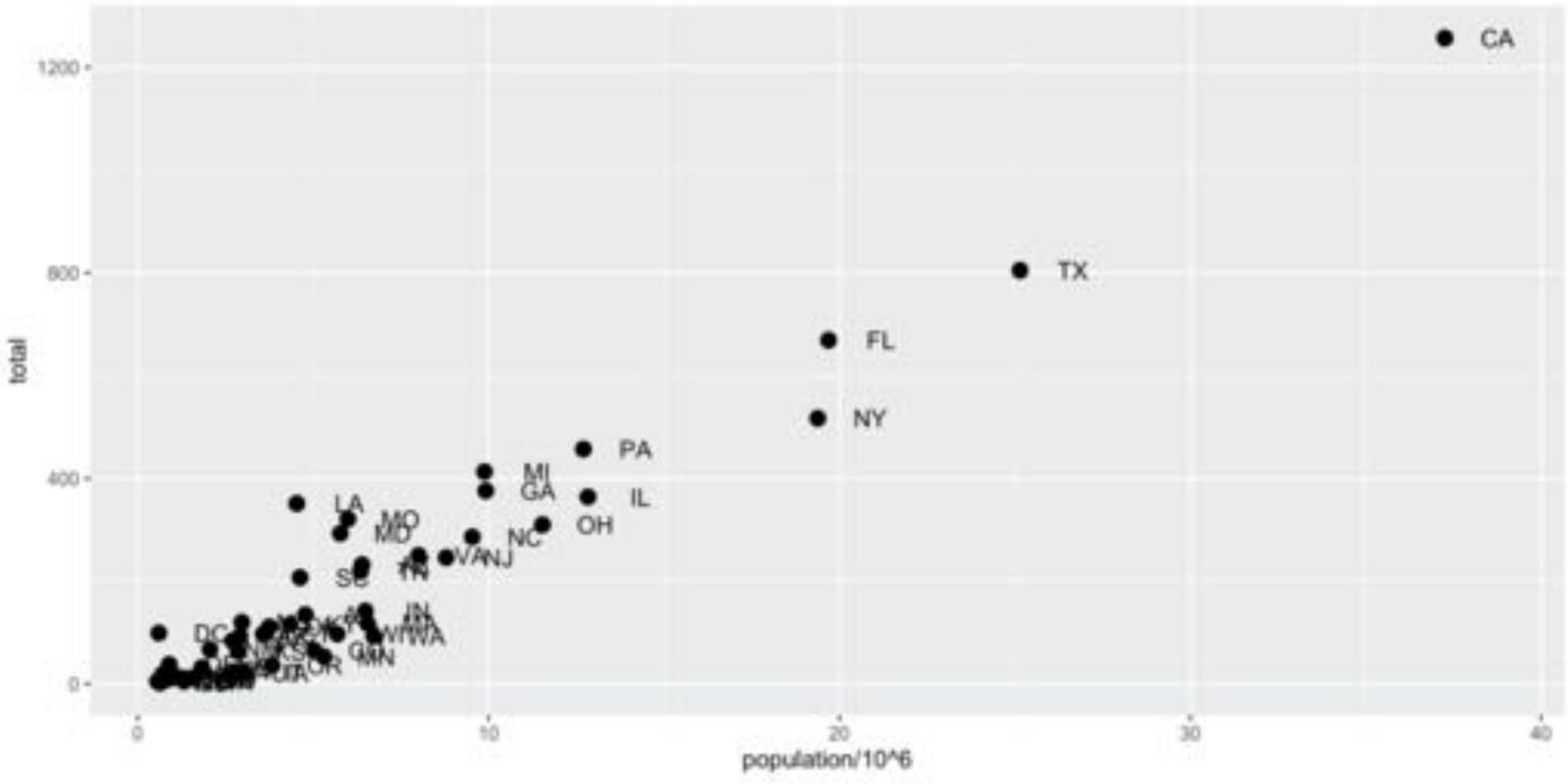
```
ggplot(data = murders) +
geom_point(aes(x = population/10^6, y = total))+
geom_text(aes(population/10^6, total, label = abb),
nudge_x = 1.5)
```



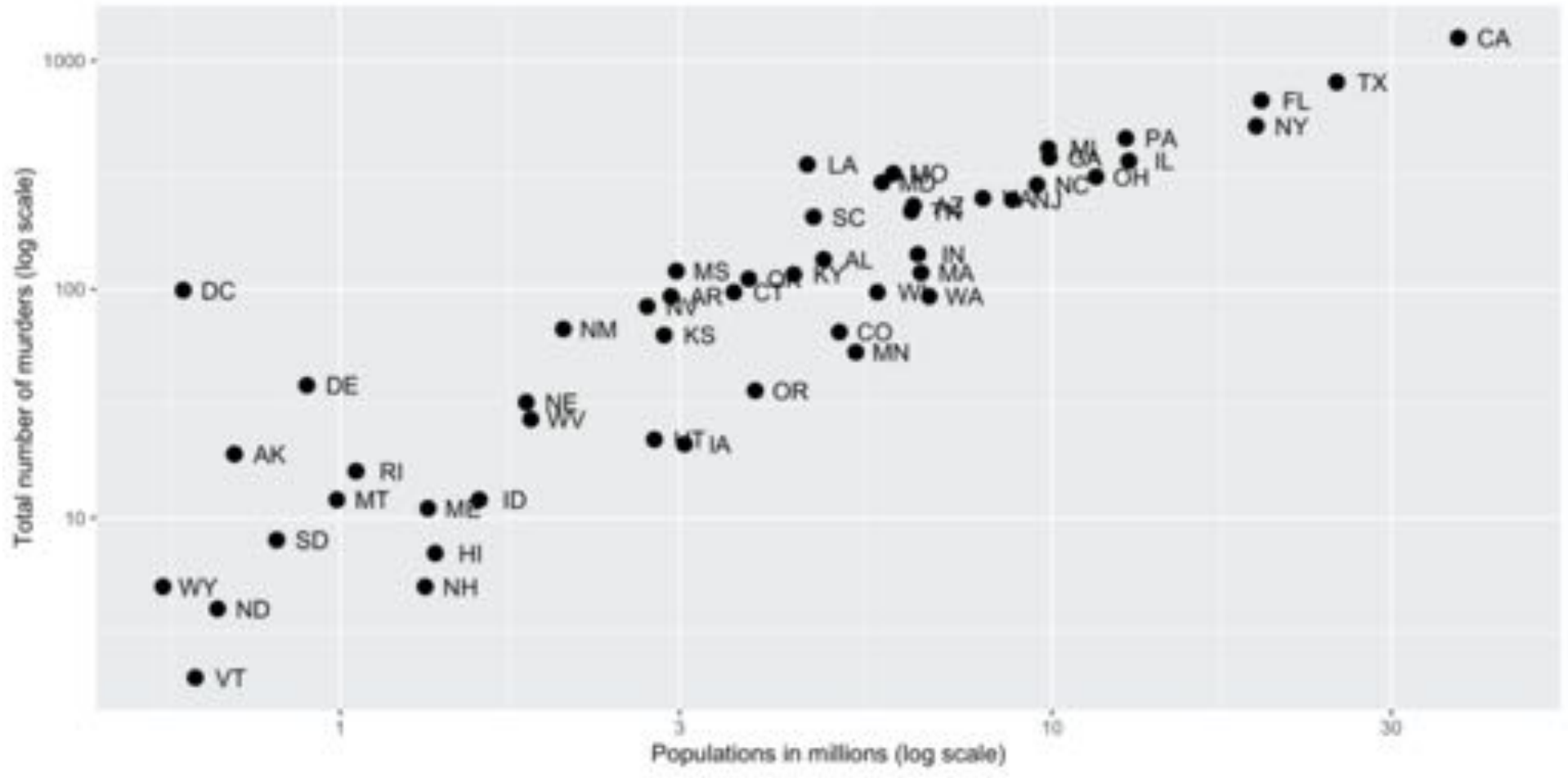
```
ggplot(murders, aes(population/10^6, total,label = abb))+
geom_point(size = 3) +
geom_text(nudge_x = 1.5)
```



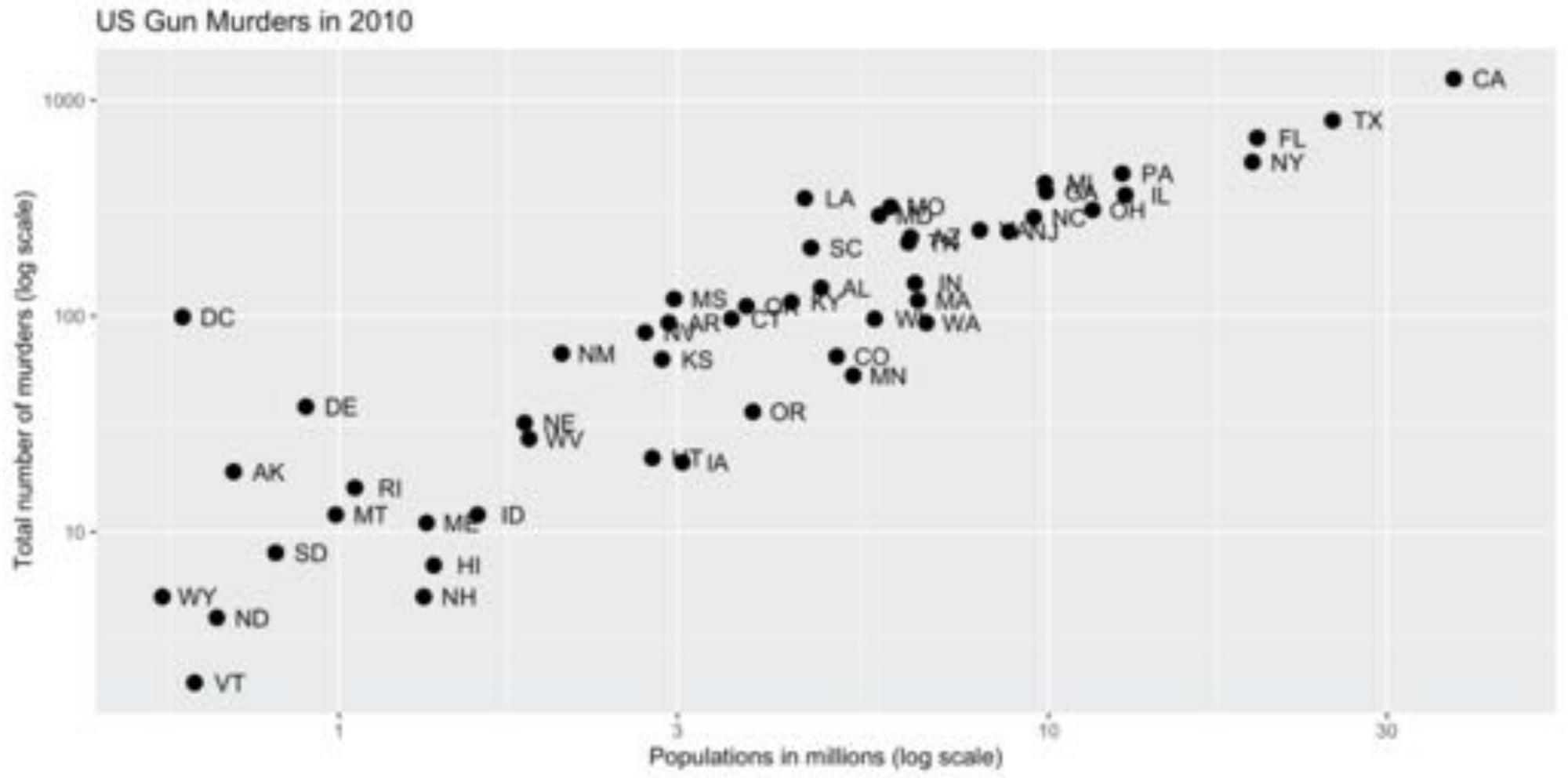
```
p= ggplot(murders,aes(population/10^6,total,label= abb))
p + geom_point(size = 3) + geom_text(nudge_x = 1.5)
```



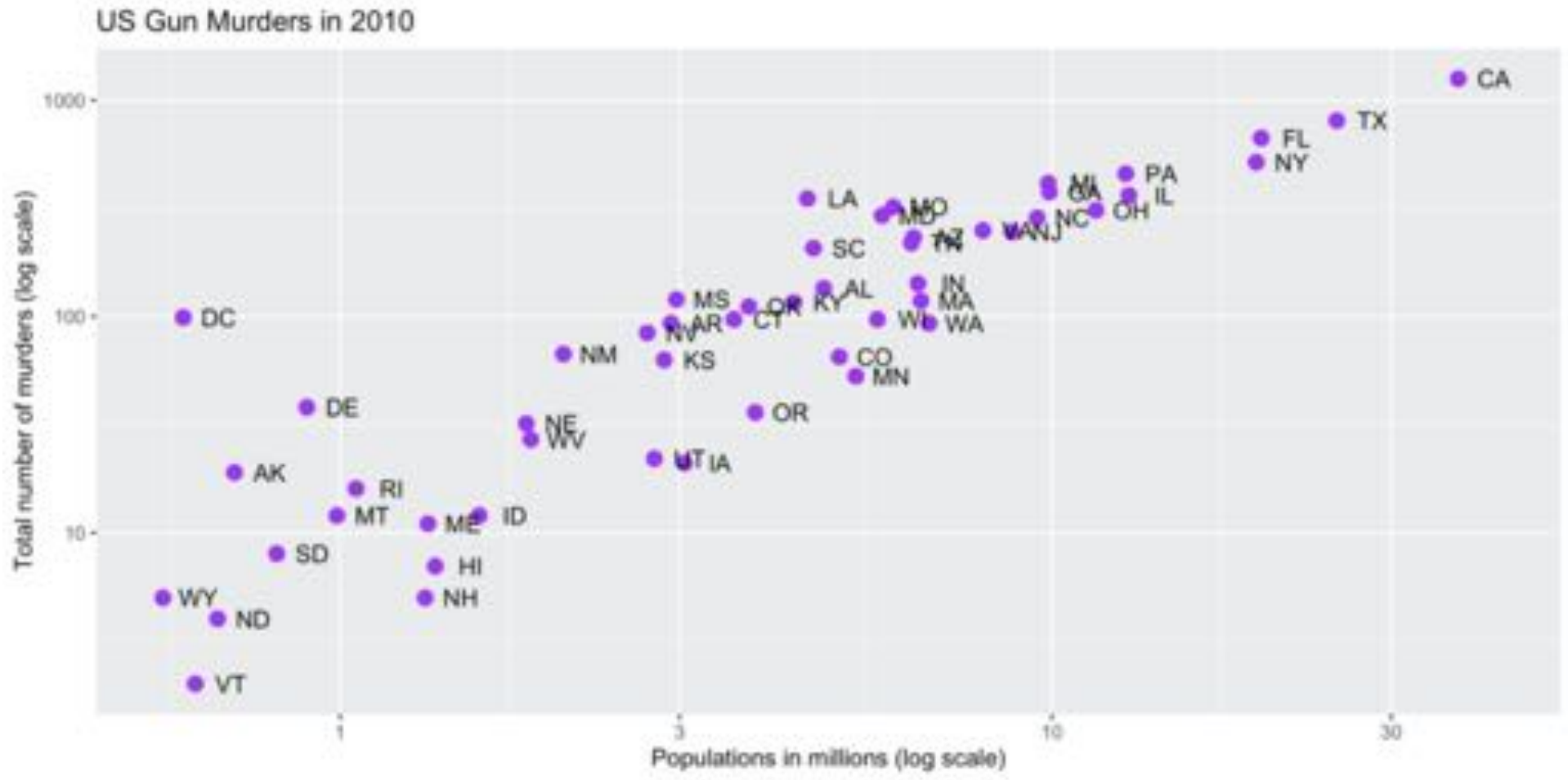
```
p + geom_point(size = 3) +
geom_text(nudge_x = 0.05) +
scale_x_continuous(trans = "log10") +
scale_y_continuous(trans = "log10")
```

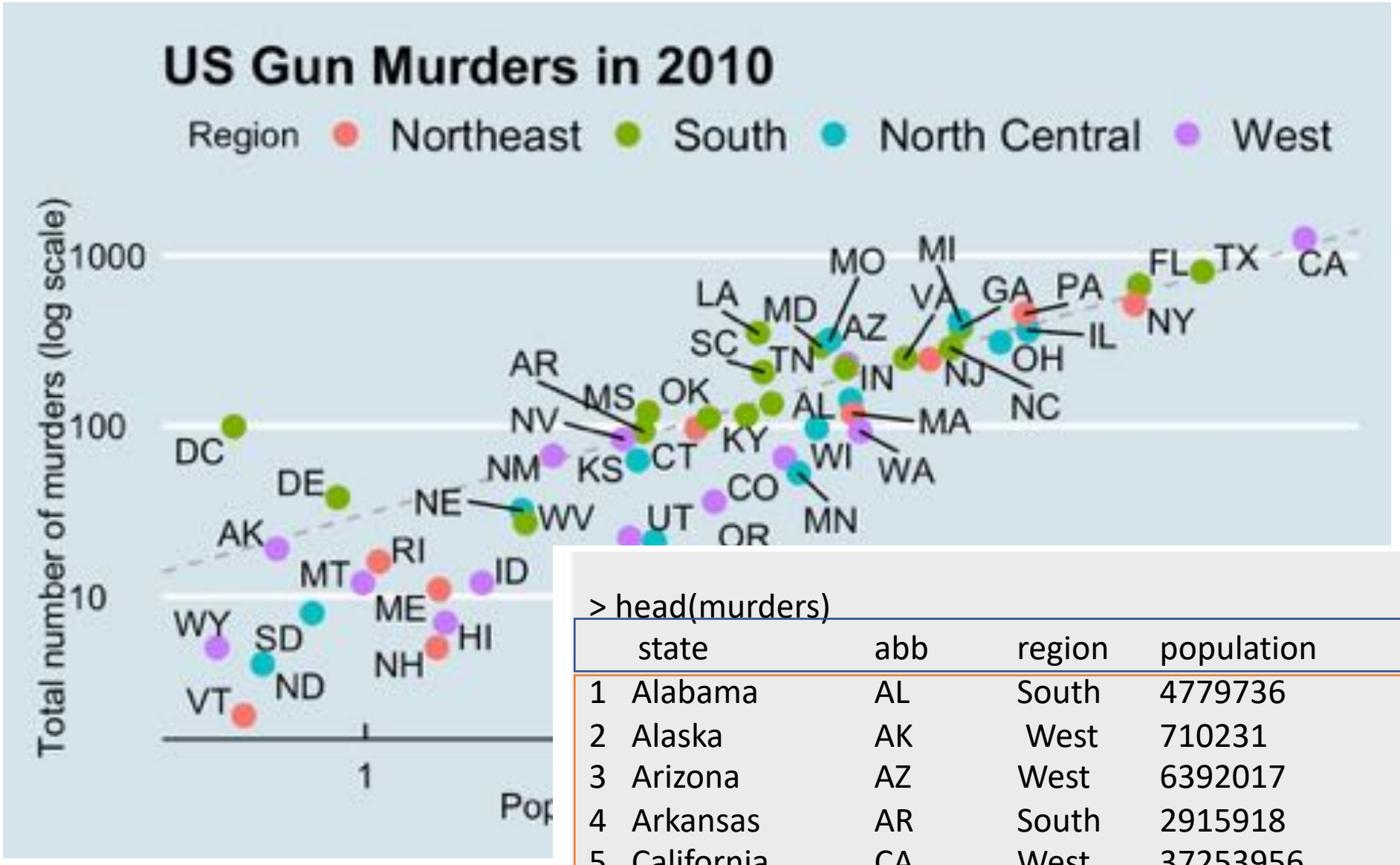


```
p + geom_point(size = 3) + geom_text(nudge_x = 0.05) +
scale_x_log10() + scale_y_log10() +
xlab("Populations in millions (log scale)") +
ylab("Total number of murders (log scale)") +
ggtitle("US Gun Murders in 2010")
```



```
p + geom_point(size = 3, color="purple") + geom_text(nudge_x = 0.05) +
scale_x_log10() + scale_y_log10() +
xlab("Populations in millions (log scale)") +
ylab("Total number of murders (log scale)") +
ggtitle("US Gun Murders in 2010")
```





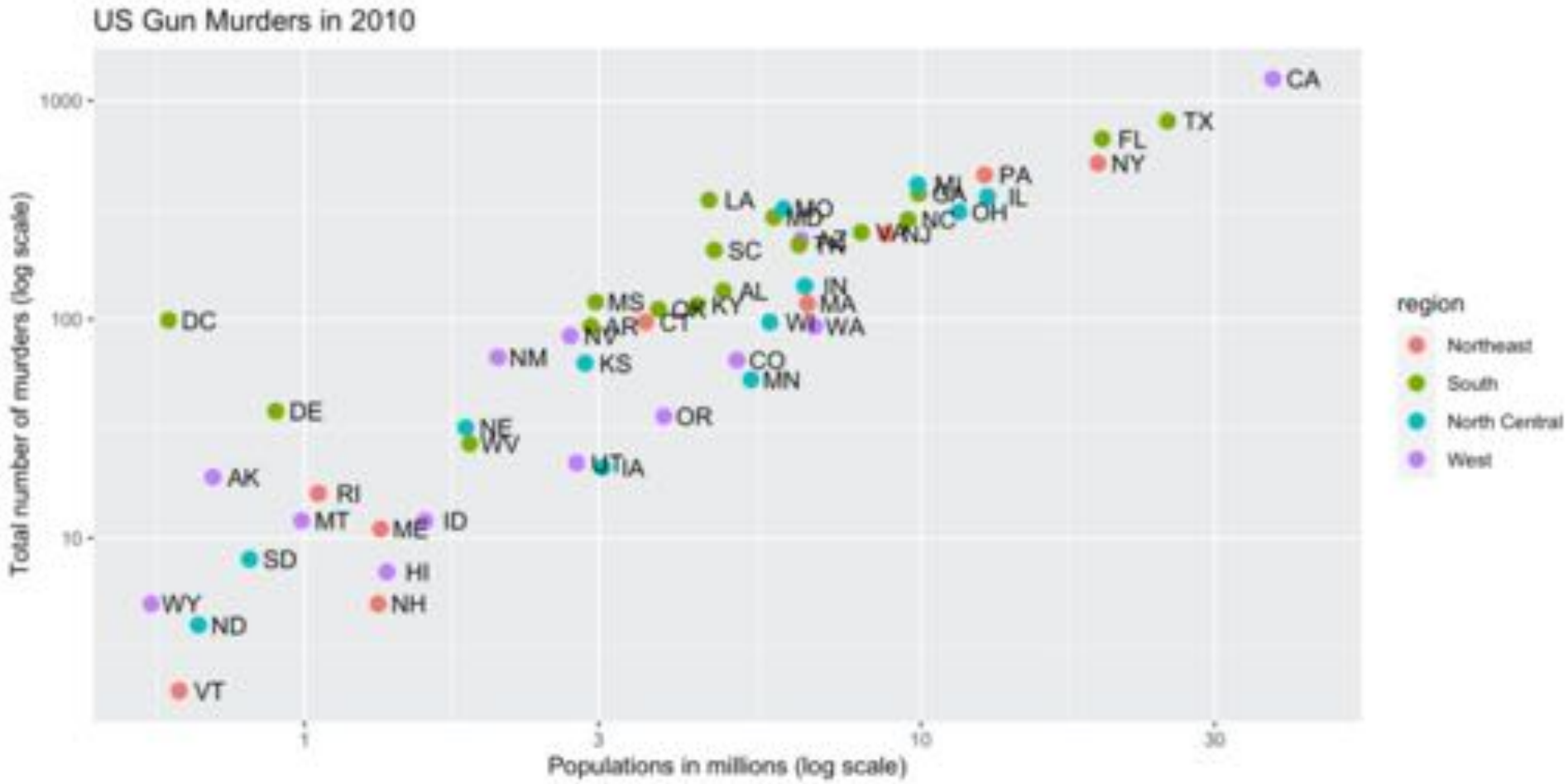
> head(murders)

	state	abb	region	population	total
1	Alabama	AL	South	4779736	135
2	Alaska	AK	West	710231	19
3	Arizona	AZ	West	6392017	232
4	Arkansas	AR	South	2915918	93
5	California	CA	West	37253956	1257
6	Colorado	CO	West	5029196	65

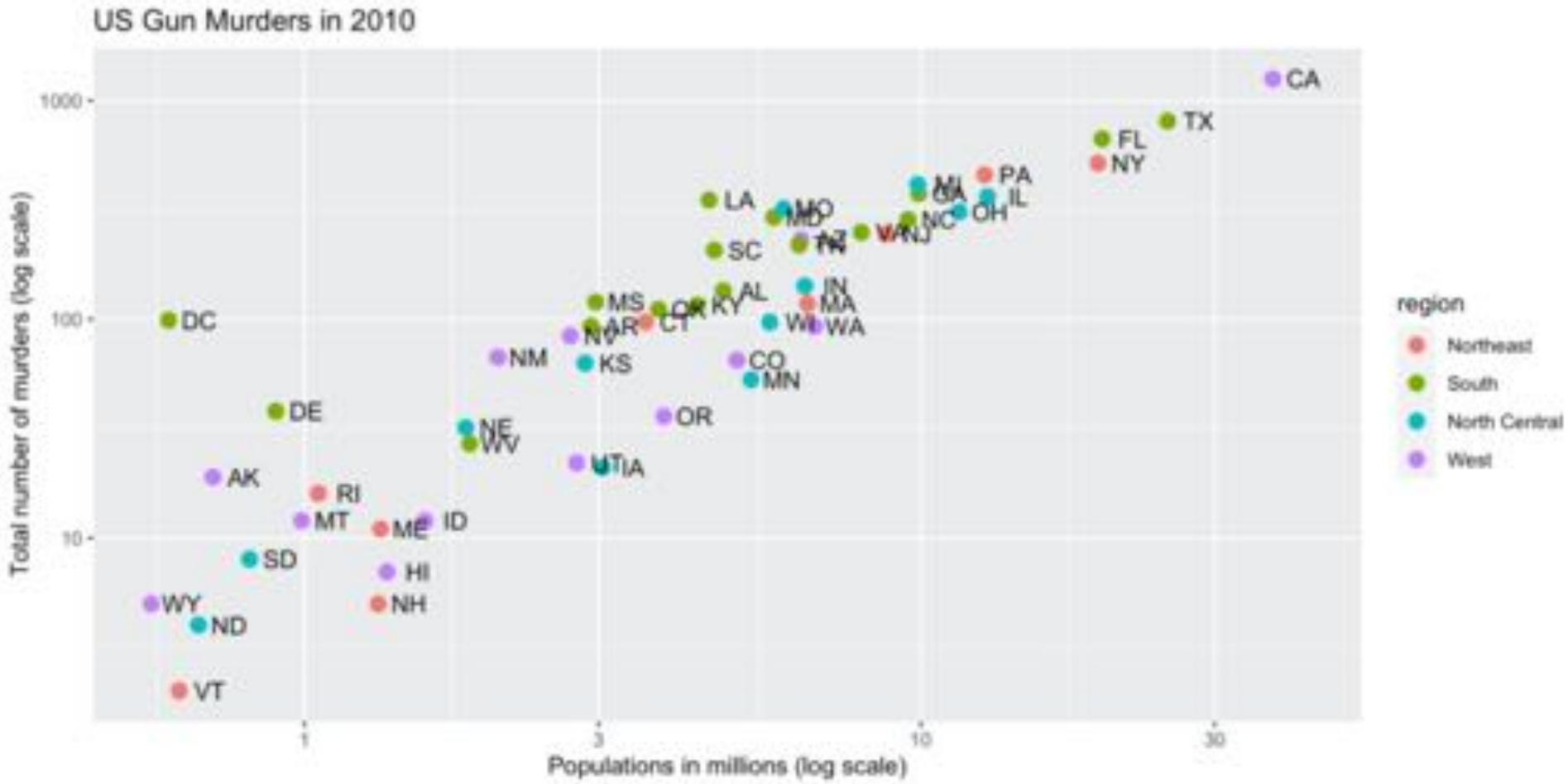
Variables

Observations

```
p + geom_point(aes(color=region), size = 3 +
geom_text(nudge_x = 0.05) +
scale_x_log10() + scale_y_log10() +
xlab("Populations in millions (log scale)") +
ylab("Total number of murders (log scale)") +
ggtitle("US Gun Murders in 2010")
```

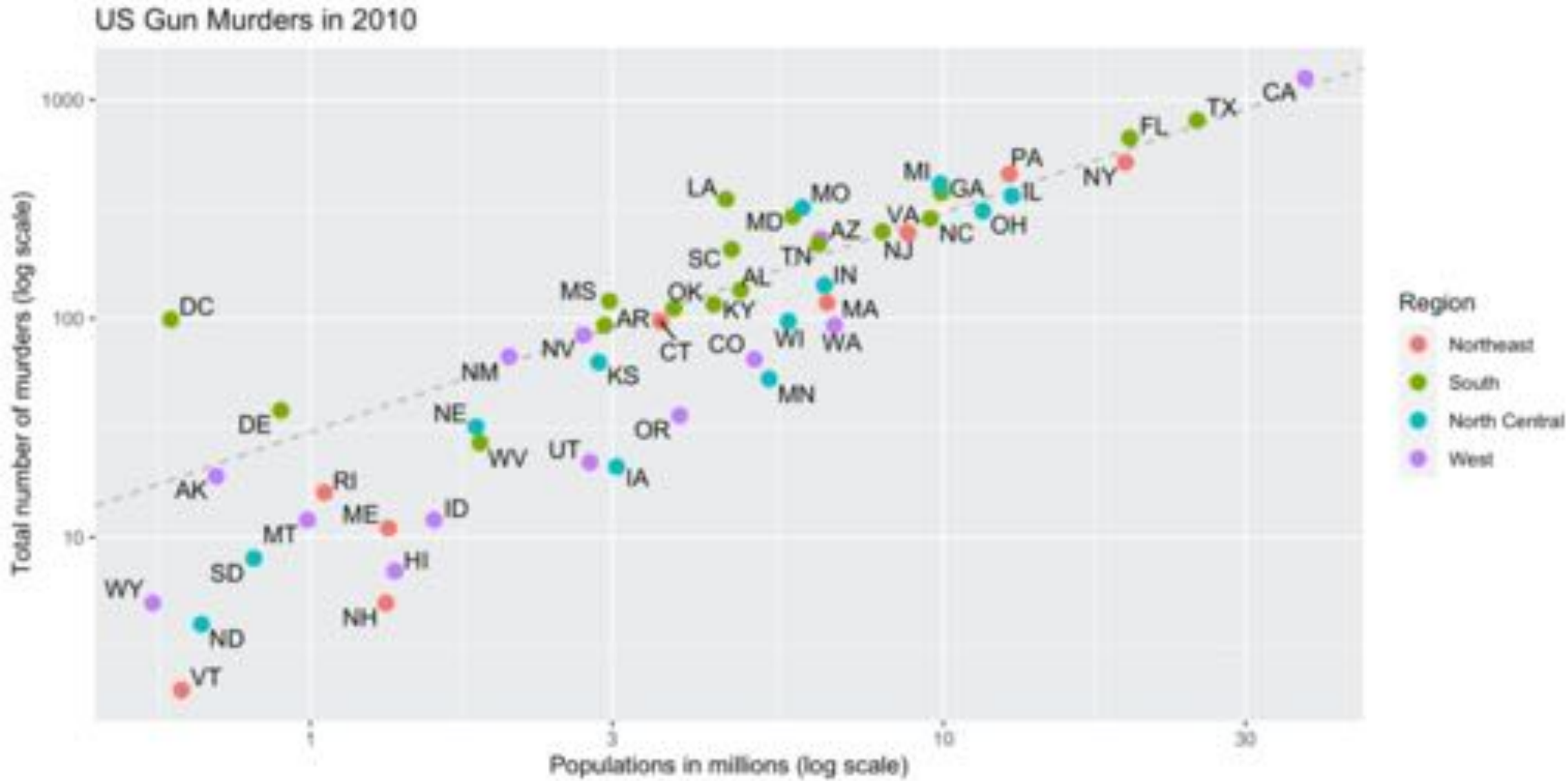


```
p + geom_point(aes(color=region), size = 3 +
geom_text(nudge_x = 0.05) +
scale_x_log10() + scale_y_log10() +
xlab("Populations in millions (log scale)") +
ylab("Total number of murders (log scale)") +
ggtitle("US Gun Murders in 2010")
```



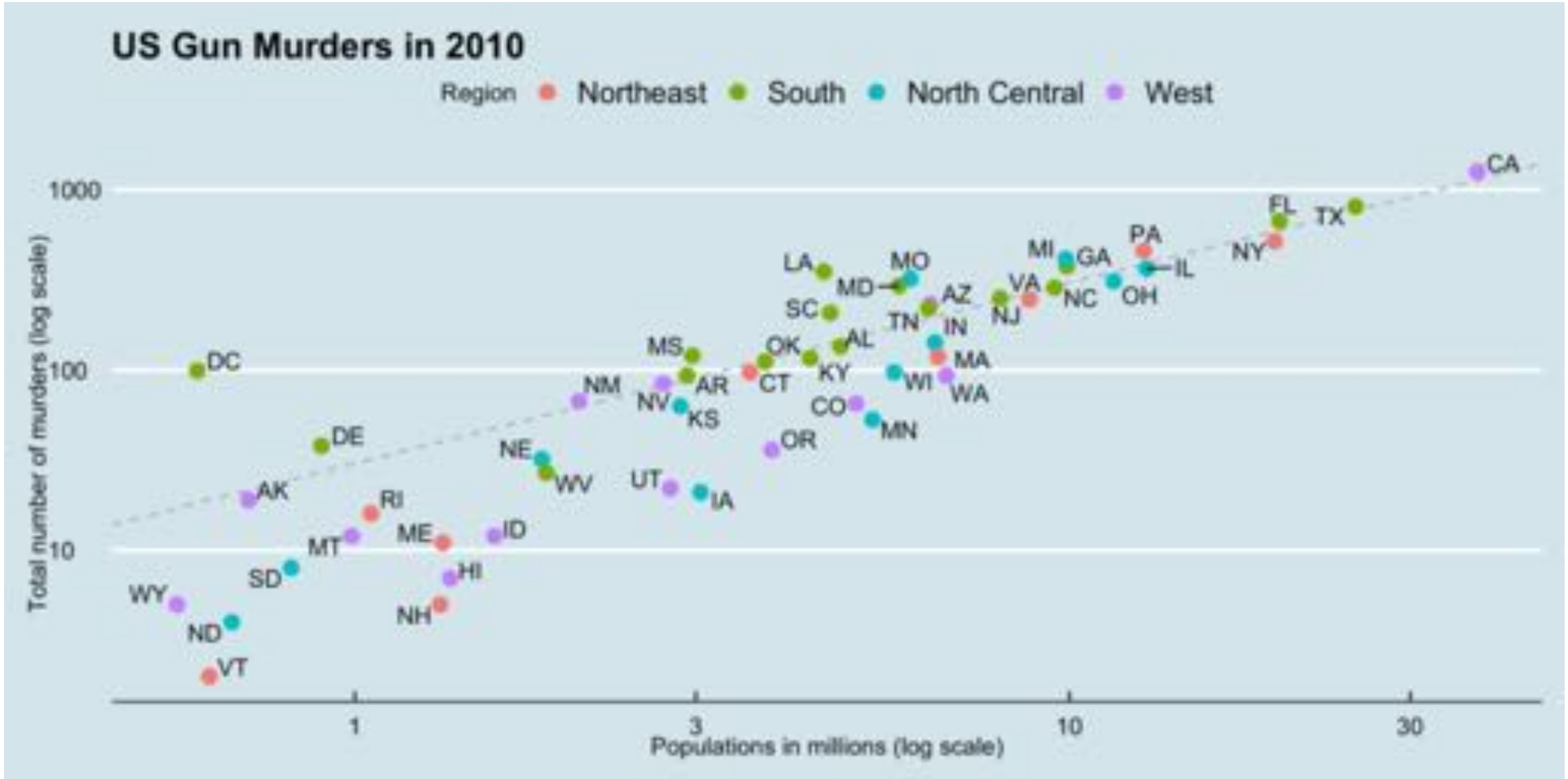
```
library(ggrepel)
r <- murders %>%
  summarize(rate = sum(total) /  sum(population) * 10^6) %>%
  pull(rate)


murders %>% ggplot(aes(population/10^6, total, label = abb)) +
geom_abline(intercept = log10(r), lty = 2, color = "darkgrey") +
geom_point(aes(col=region), size = 3) +
geom_text_repel() +
scale_x_log10() +
scale_y_log10() +
xlab("Populations in millions (log scale)") +
ylab("Total number of murders (log scale)") +
ggtitle("US Gun Murders in 2010") +
scale_color_discrete(name = "Region")
```


US Gun Murders in 2010
Total number of murders (log scale)
1000
100
10
Populations in millions (log scale)
1
3
10
30
CA
TX
FL
NY
PA
IL
MI
GA
OH
NC
VA
NJ
AZ
TN
MO
MD
LA
SC
AL
IN
MA
WA
WI
MN
CO
KY
OK
MS
AR
CT
NV
NM
KS
NE
WV
OR
UT
IA
DE
DC
AK
RI
ME
ID
MT
SD
HI
WY
NH
ND
VT
Region
Northeast
South
North Central
West

```
library(ggthemes)
r <- murders %>%
  summarize(rate = sum(total) /  sum(population) * 10^6) %>%
  pull(rate)


ggplot(murders, aes(population/10^6, total, label = abb)) +
geom_abline(intercept = log10(r), lty = 2, color = "darkgrey") +
geom_point(aes(col=region), size = 3) +
geom_text_repel() +
scale_x_log10() +
scale_y_log10() +
xlab("Populations in millions (log scale)") +
ylab("Total number of murders (log scale)") +
ggtitle("US Gun Murders in 2010") +
scale_color_discrete(name = "Region") +
theme_economist()
```


US Gun Murders in 2010
Region
Northeast
South
North Central
West
Total number of murders (log scale)
1000
100
10
Populations in millions (log scale)
1
3
10
30
CA
TX
FL
NY
PA
IL
MI
GA
OH
NC
VA
NJ
AZ
TN
IN
MO
MD
LA
SC
AL
OK
KY
MS
AR
CT
MA
WI
WA
NM
NV
KS
CO
MN
OR
NE
WV
UT
IA
DC
DE
AK
RI
ME
MT
ID
SD
HI
NH
WY
ND
VT

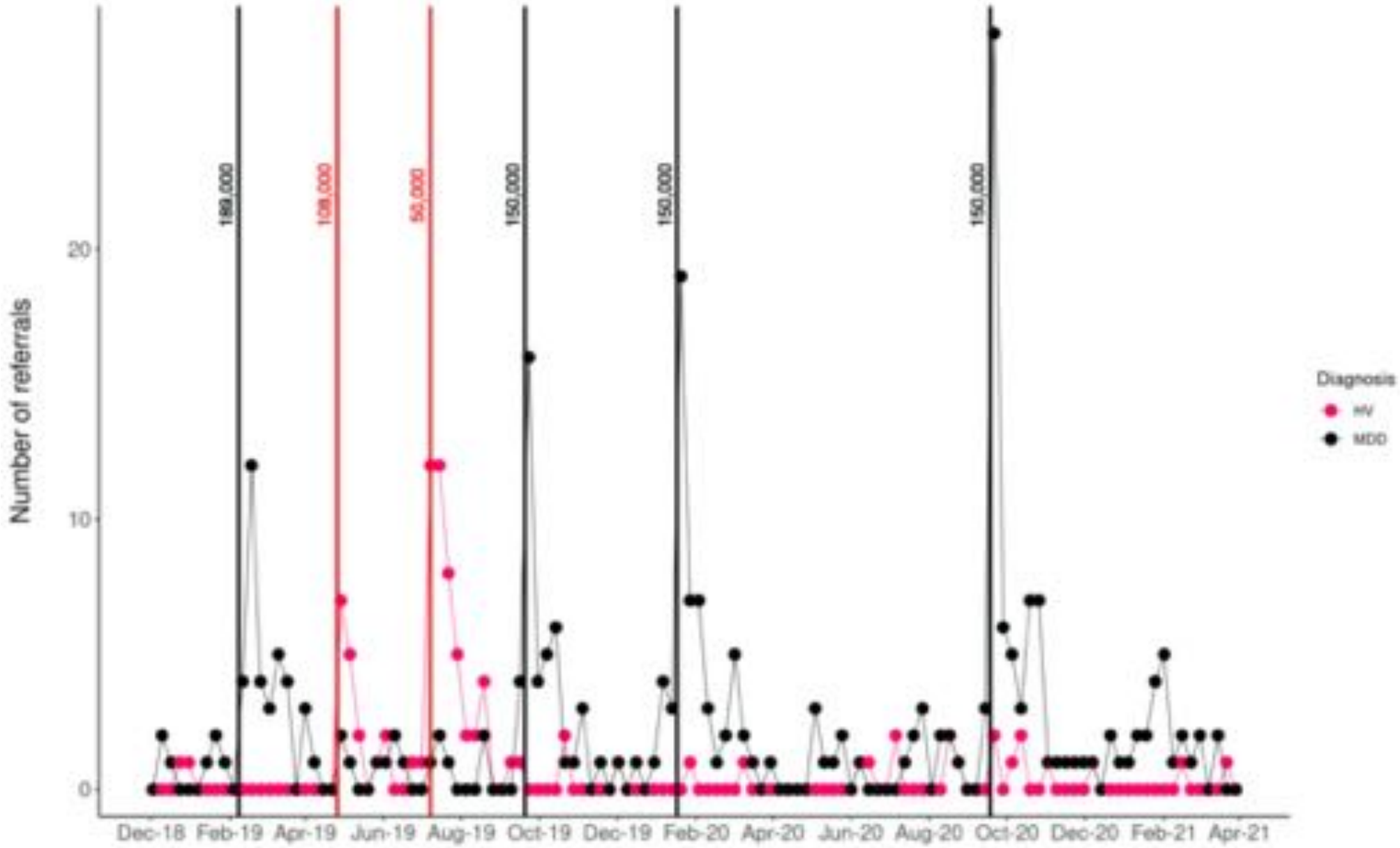
Characterization and Treatment of Depression

CAT-D

Characterization Study (*all participants*)
Behavior, Brains, Biology

- Assess behavioral changes using interviews and questionnaires
  - Interviews = KSADS
  - Questionnaires = DAWBA to screen, MFQ, SCARED, others
- Assess brain changes using fMRI and MEG
- Assess biological changes using blood samples

2 Treatment Studies (*subgroup of participants*): Outpatient and Inpatient

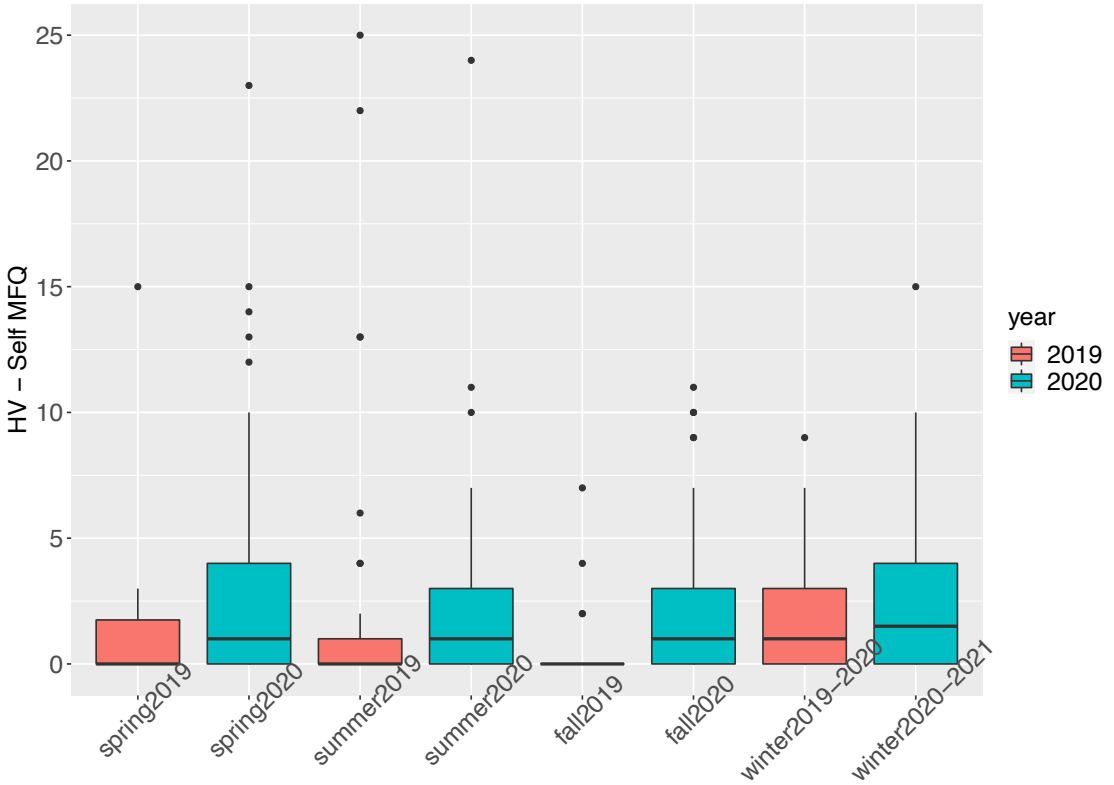

Number of referrals
0
10
20
189,000
108,000
50,000
150,000
150,000
150,000
Dec-18
Feb-19
Apr-19
Jun-19
Aug-19
Oct-19
Dec-19
Feb-20
Apr-20
Jun-20
Aug-20
Oct-20
Dec-20
Feb-21
Apr-21
Diagnosis
HIV
MDD

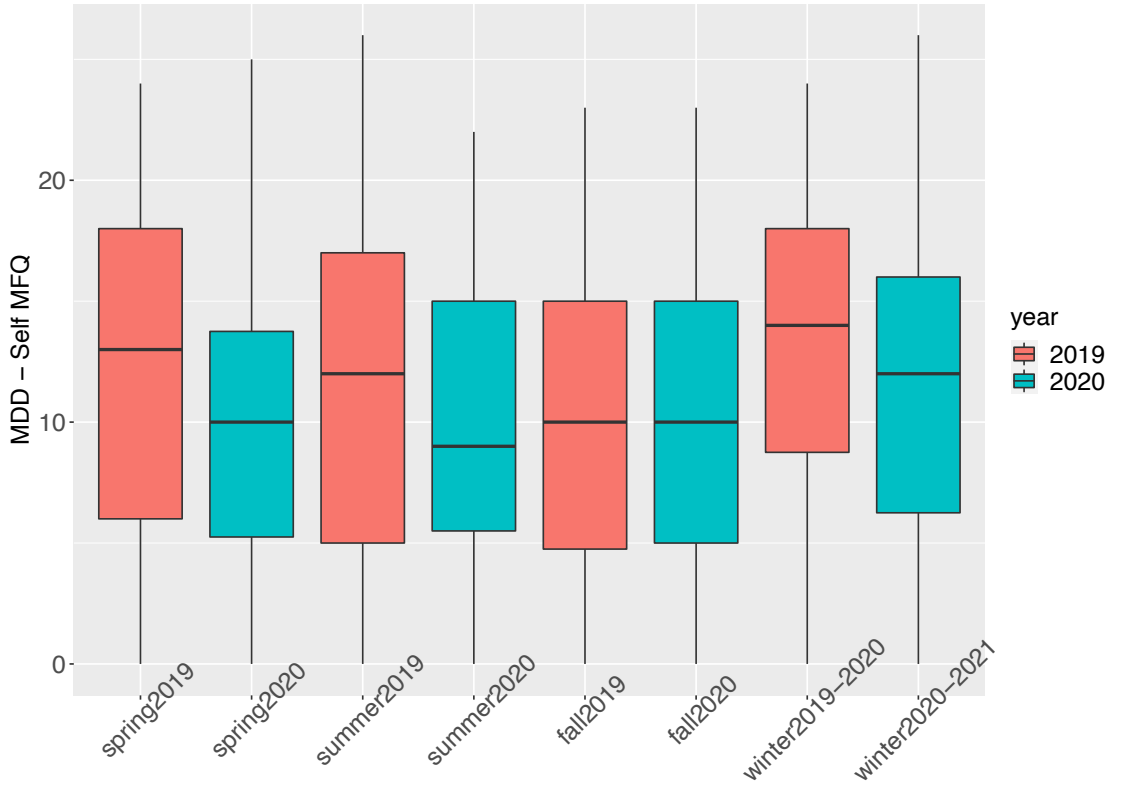
Participant Demographics

N = 194	**M(SD) | N(%)**
Age	16.11 (1.59)
Sex	
Male	64 (33%)
Female	130 (67%)
Diagnostic Group	
HV	85 (56%)
MDD	109 (44%)

HV = healthy volunteer; MDD = volunteer with major depressive disorder

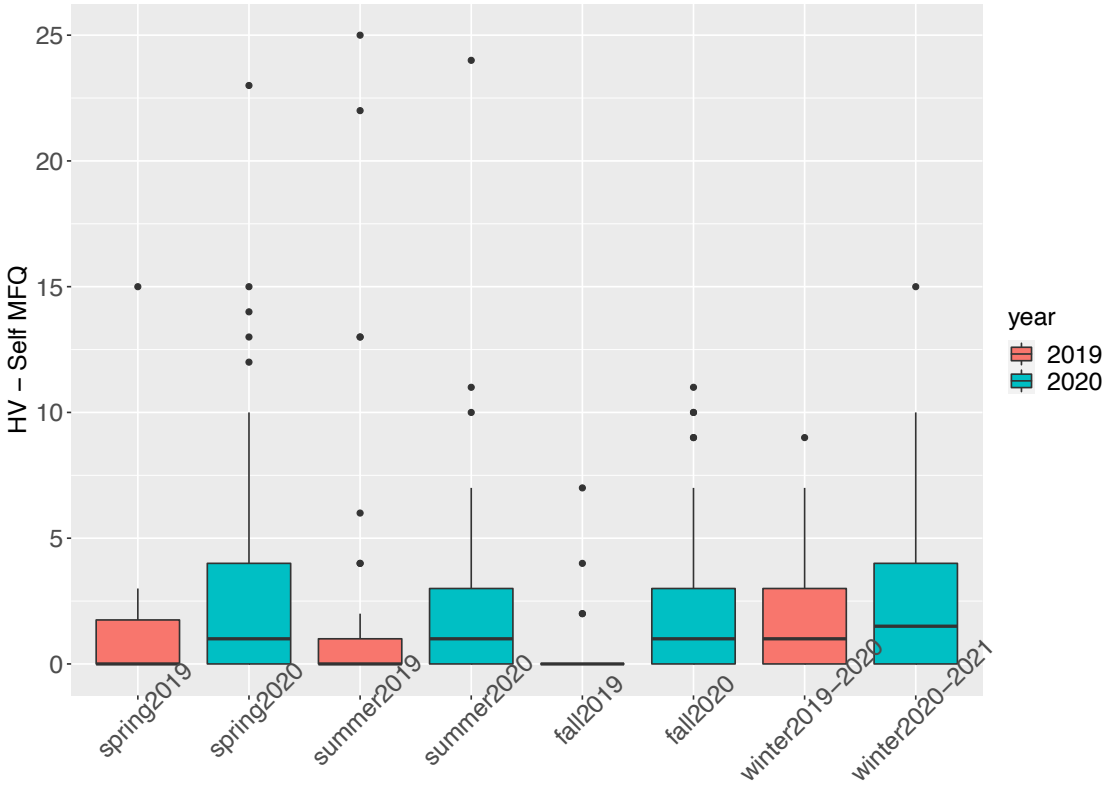
```
 covidData %>% filter(Diagnosis=="HV") %>%
    ggplot(aes(x=season, y=mfq, fill=year)) + geom_boxplot() +
    labs(x="", y=paste("HV -", measureLabel))+
    theme(axis.text.x = element_text(angle =  45),
axis.text=element_text(size=16),axis.title=element_text(size=16),
    legend.title = element_text(size = 16), legend.text = element_text(size = 16))
```

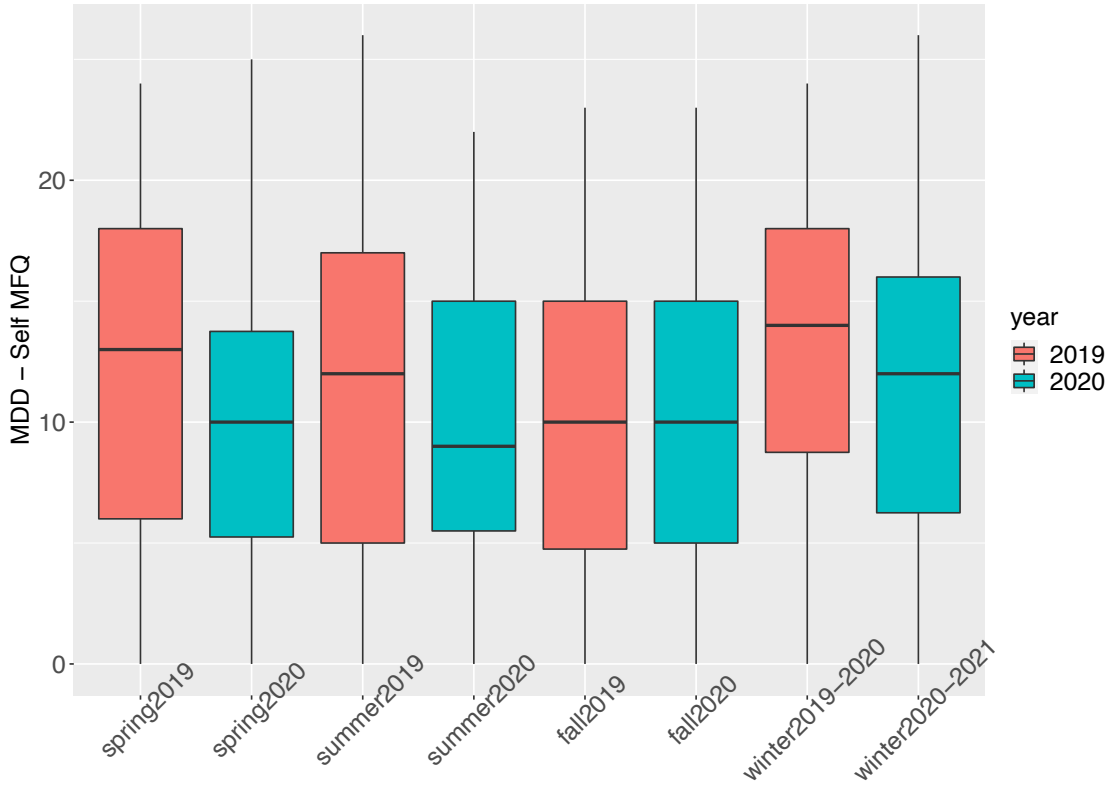




Preliminary results

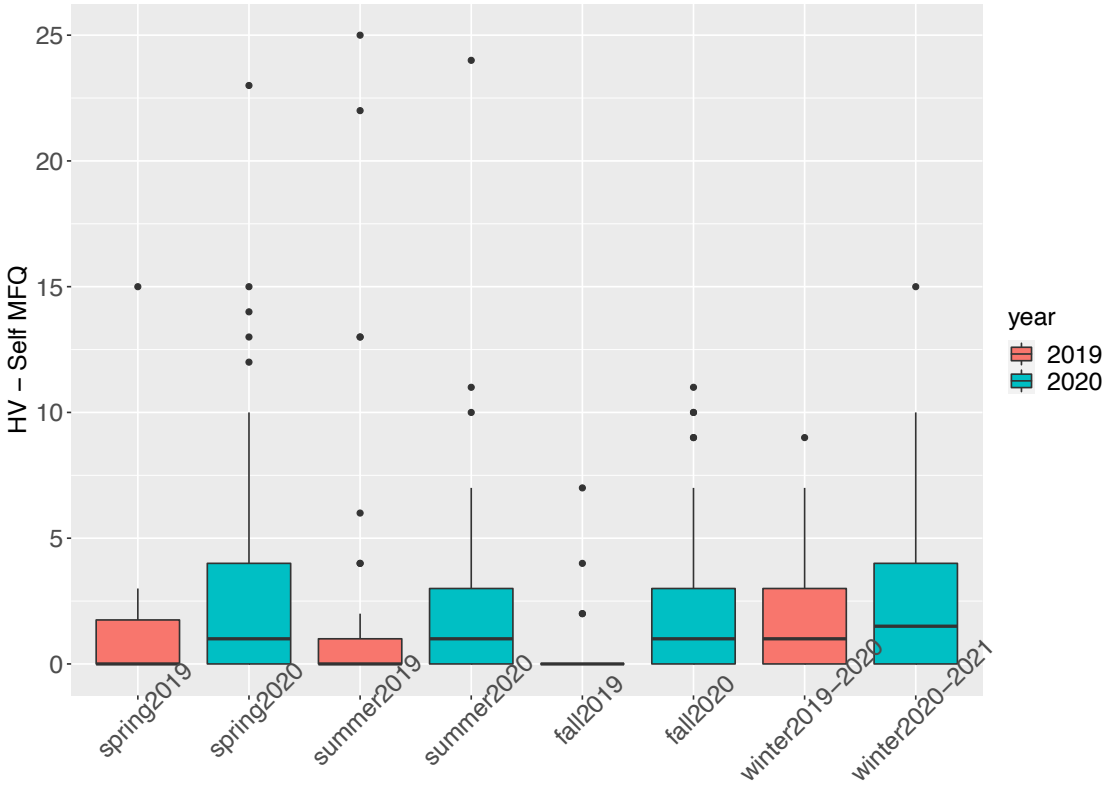
```
 covidData %>% filter(Diagnosis=="HV") %>%
    ggplot(aes(x=season, y=mfq, fill=year)) + geom_boxplot() +
    labs(x="", y=paste("HV -", measureLabel))+
    theme(axis.text.x = element_text(angle =  45),
axis.text=element_text(size=16),axis.title=element_text(size=16),
    legend.title = element_text(size = 16), legend.text = element_text(size = 16))
```

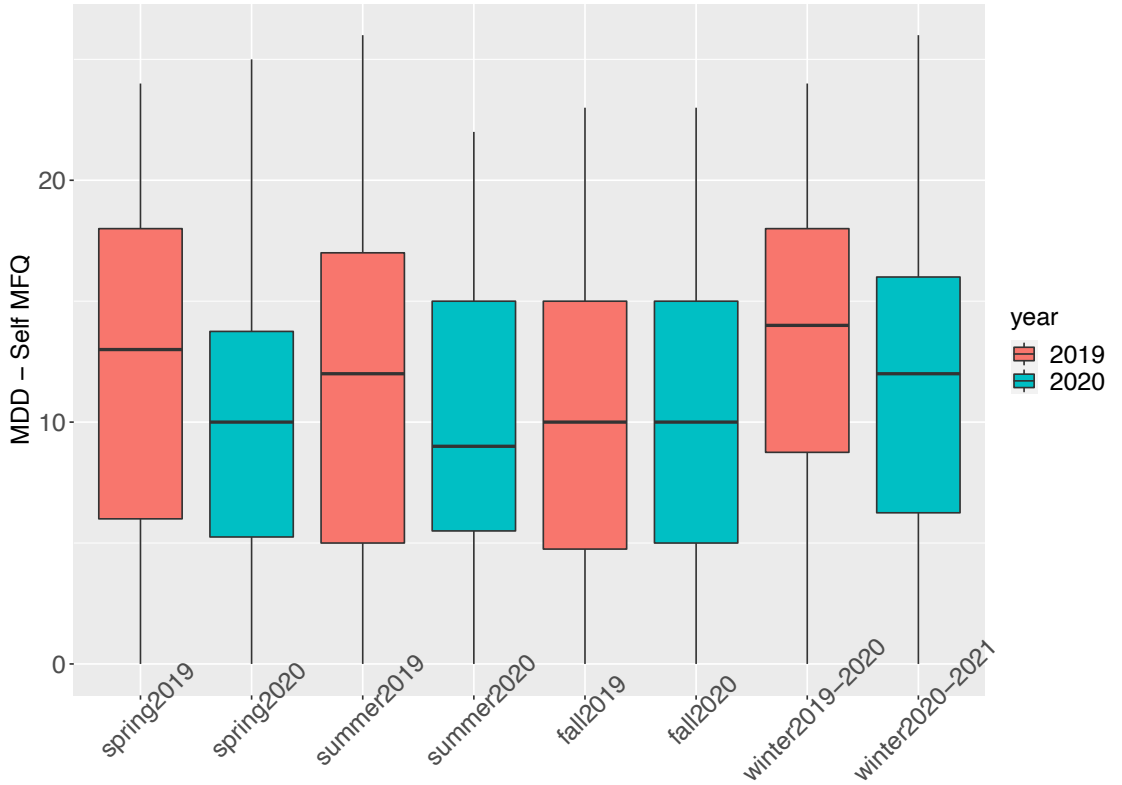




Preliminary results

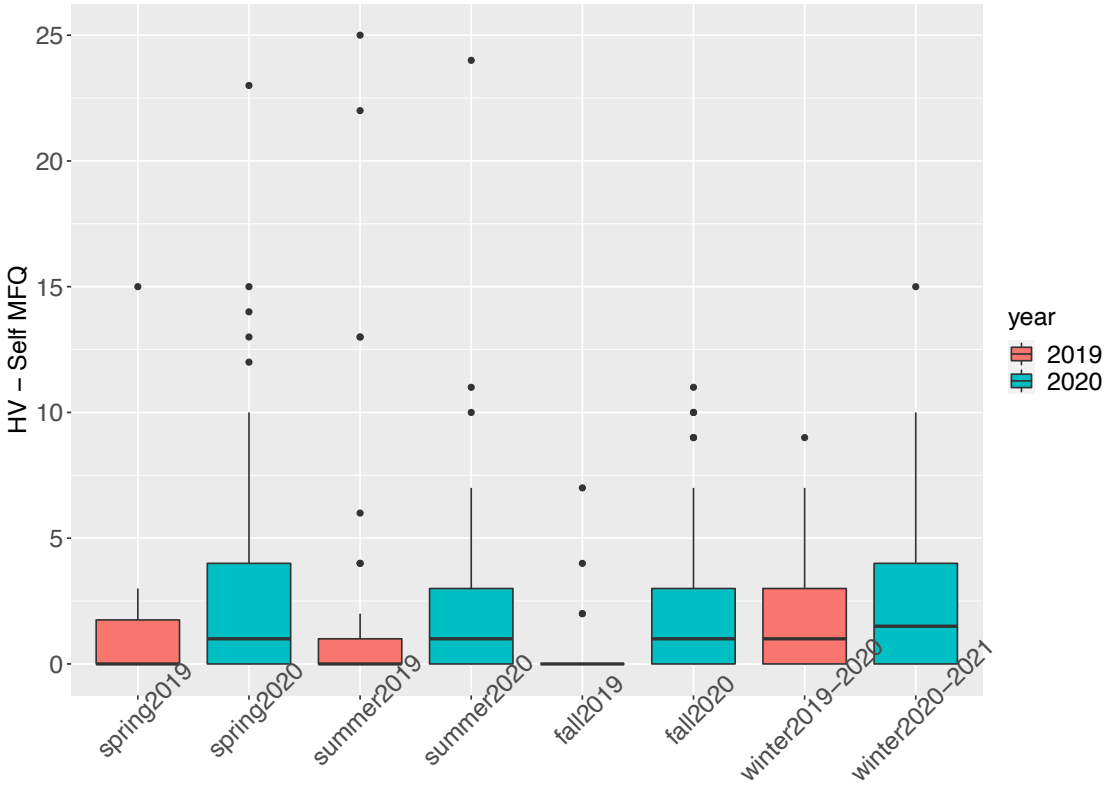
```
 covidData %>% filter(Diagnosis=="HV") %>%
    ggplot(aes(x=season, y=mfq, fill=year)) + geom_boxplot() +
    labs(x="", y=paste("HV -", measureLabel))+
    theme(axis.text.x = element_text(angle =  45),
axis.text=element_text(size=16),axis.title=element_text(size=16),
    legend.title = element_text(size = 16), legend.text = element_text(size = 16))
```

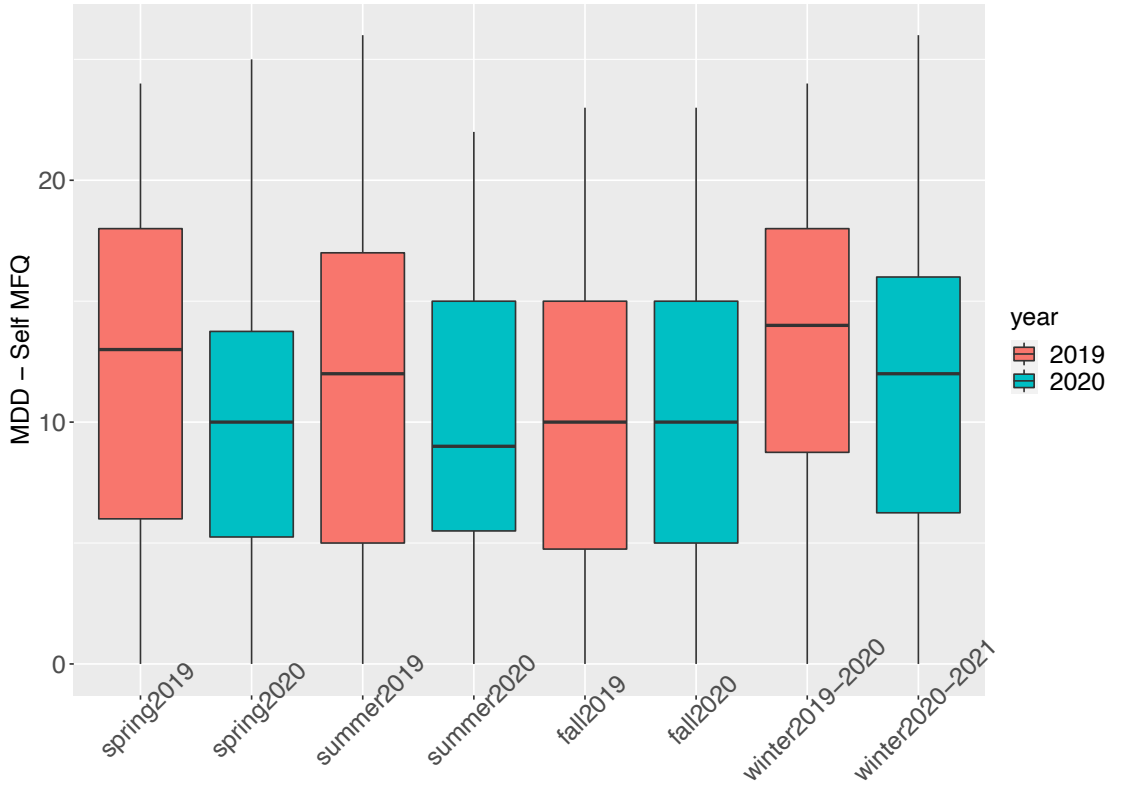




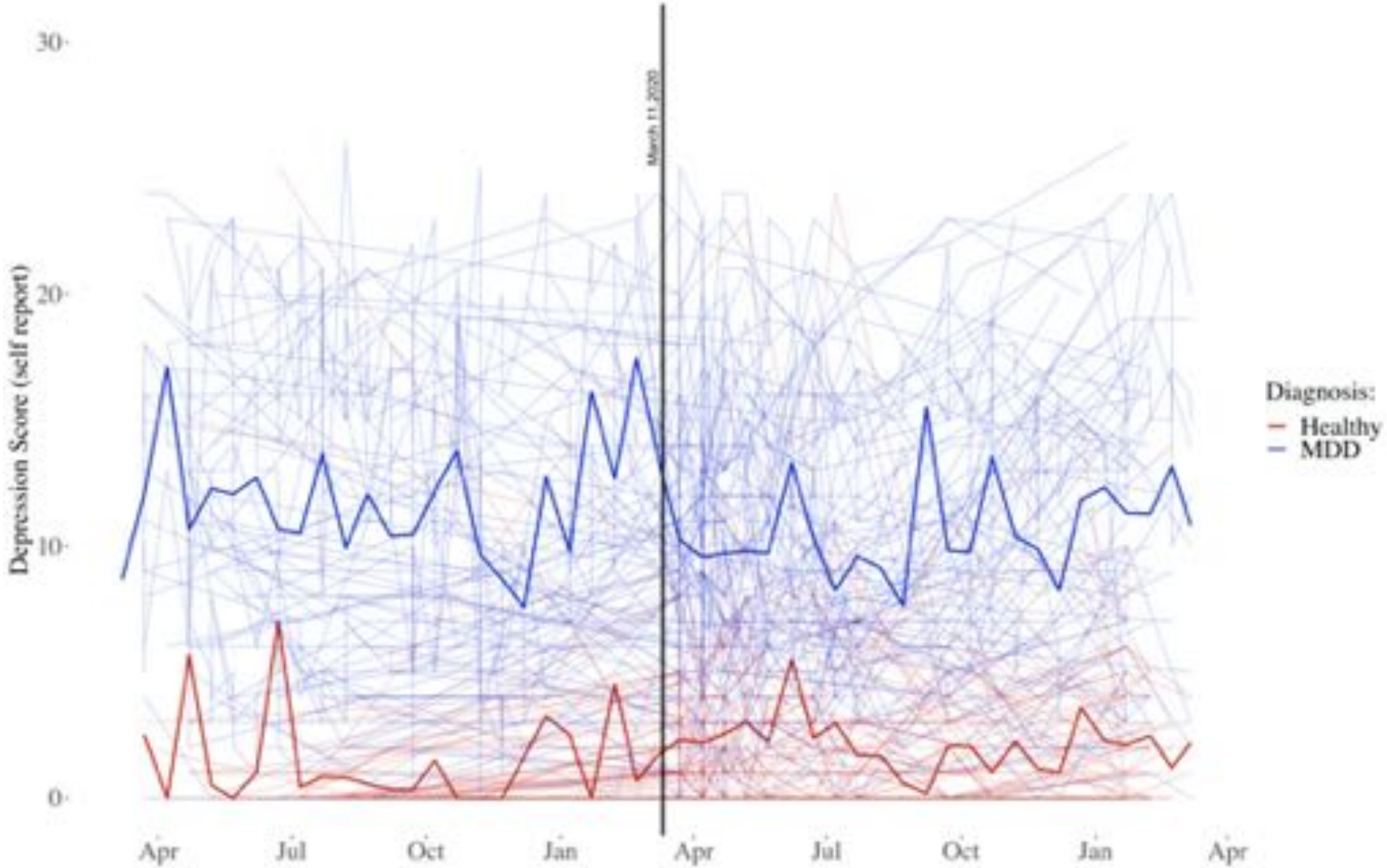
Preliminary results

```
 covidData %>% filter(Diagnosis=="HV") %>%
    ggplot(aes(x=season, y=mfq, fill=year)) + geom_boxplot() +
    labs(x="", y=paste("HV -", measureLabel))+
    theme(axis.text.x = element_text(angle =  45),
axis.text=element_text(size=16),axis.title=element_text(size=16),
    legend.title = element_text(size = 16), legend.text = element_text(size = 16))
```



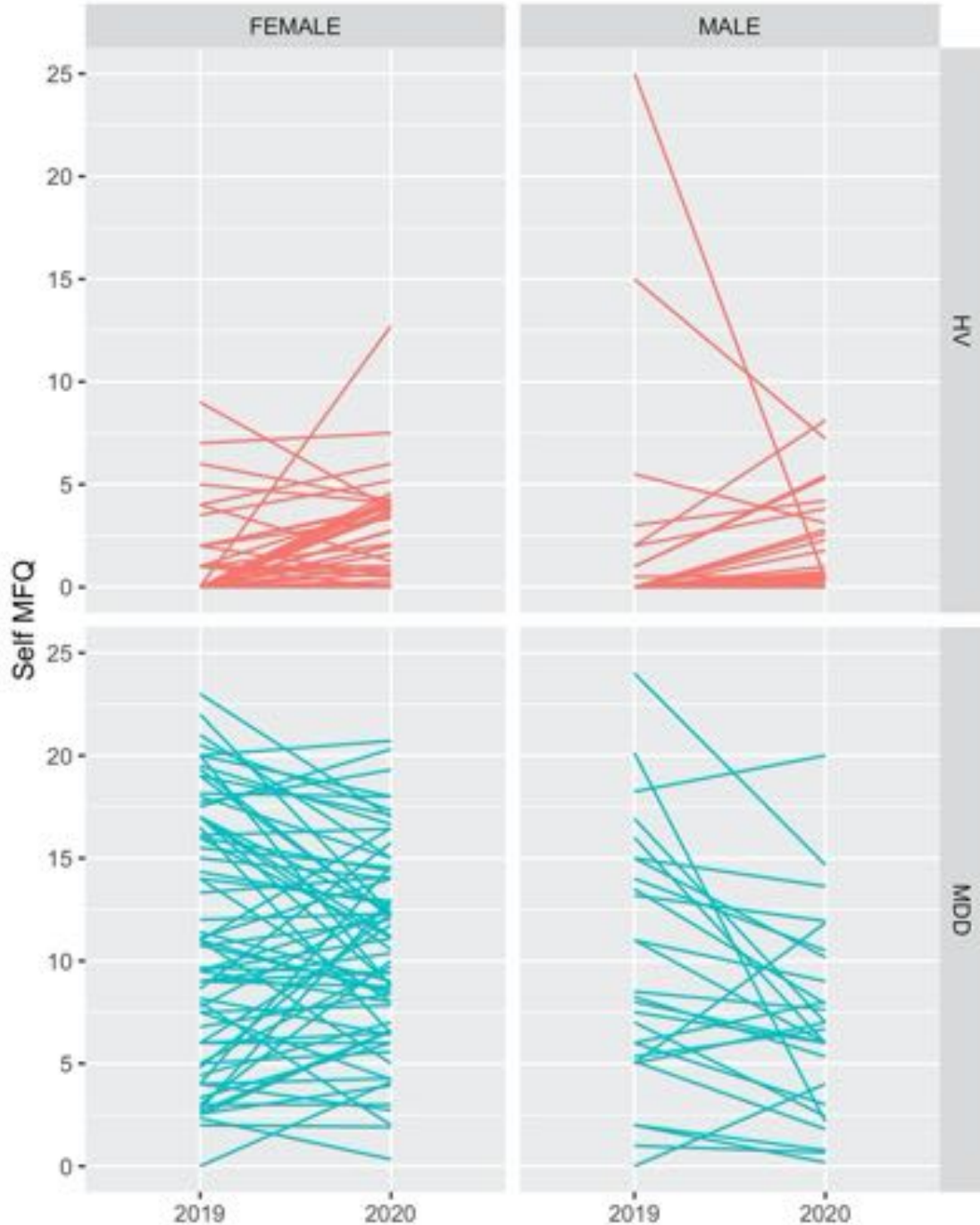


Preliminary results



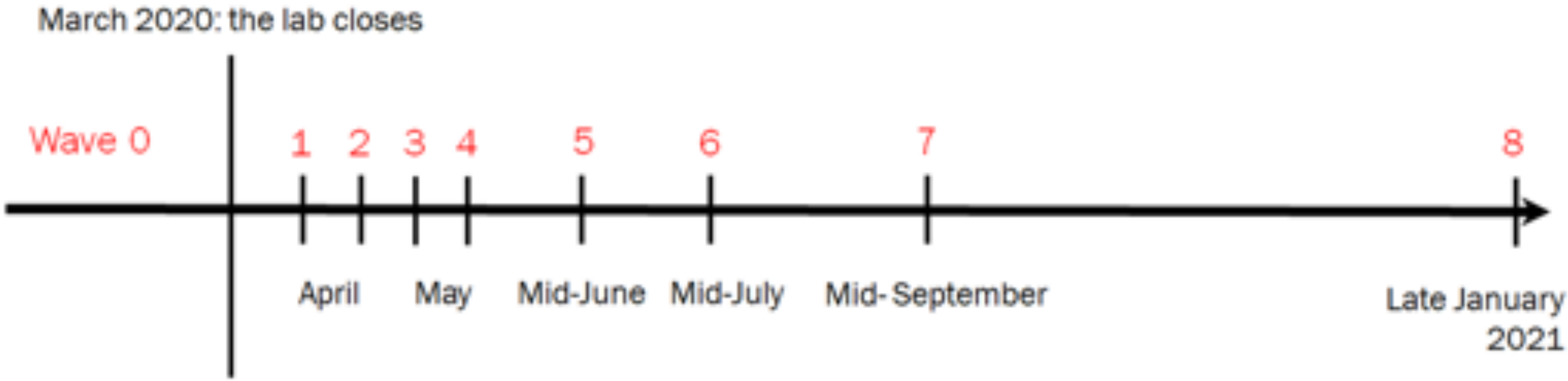
Preliminary results

Checking for outliers



Preliminary results

CoRonavIruS Impact Survey (CRISIS)

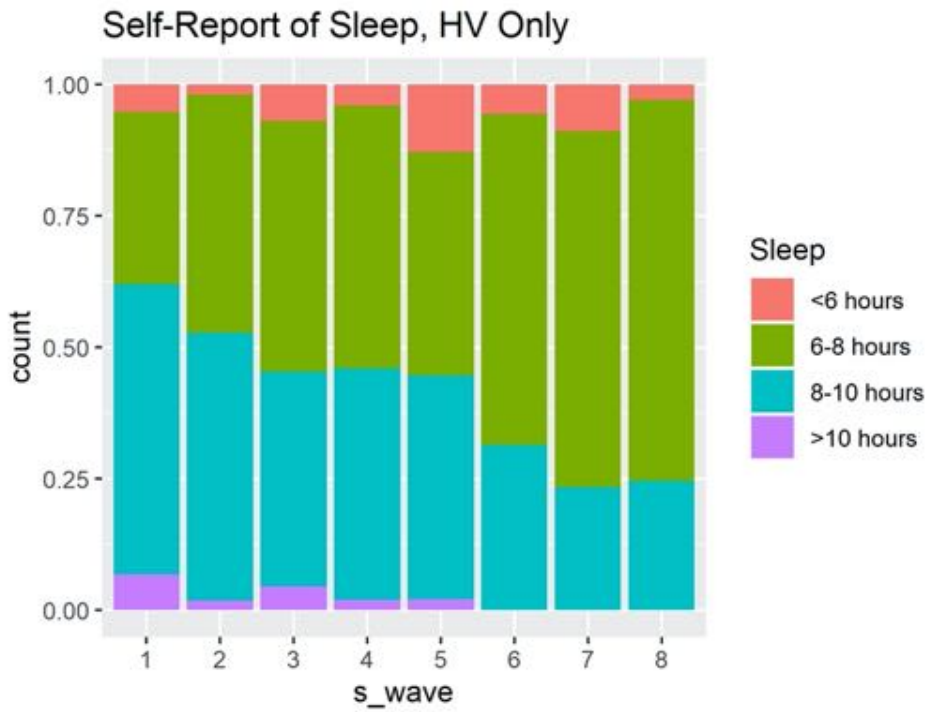


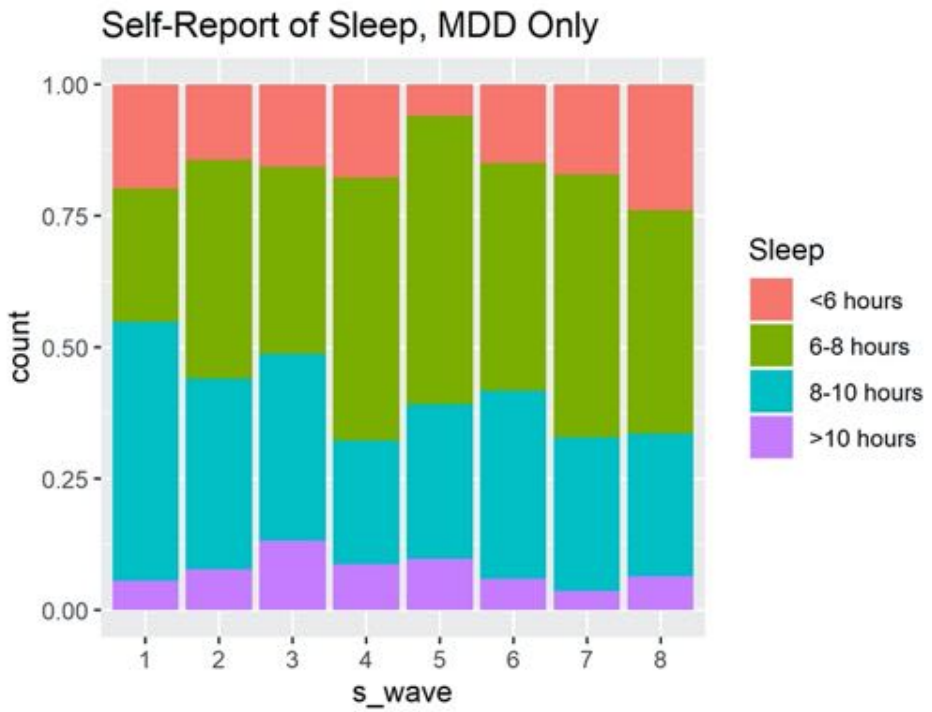
MFQ, SCARED

August 2017 through March 17, 2020

Preliminary results

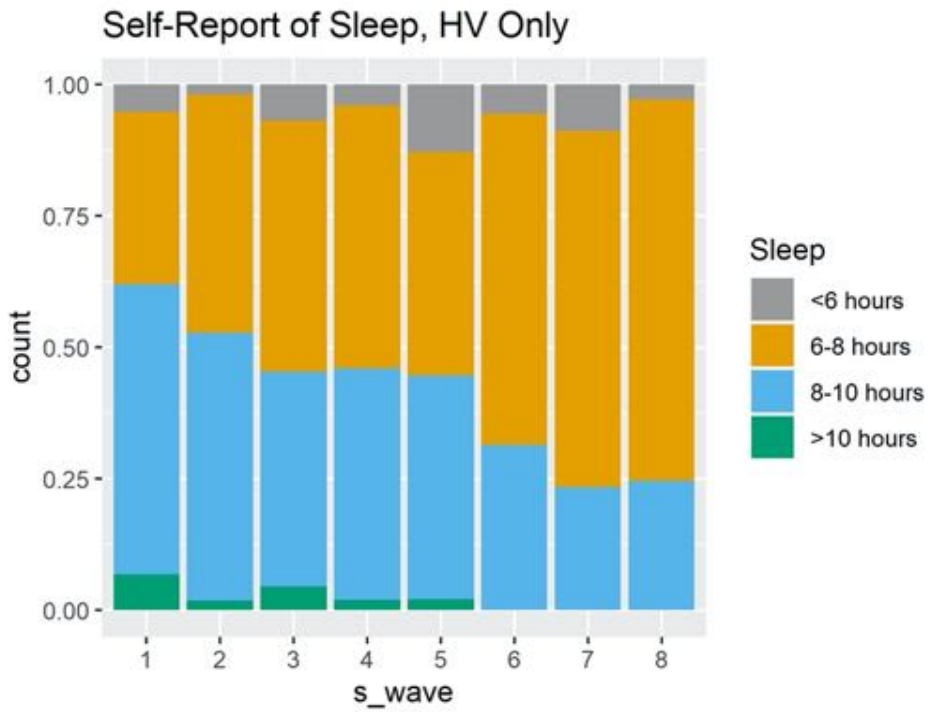
Sleep pattern during COVID-19

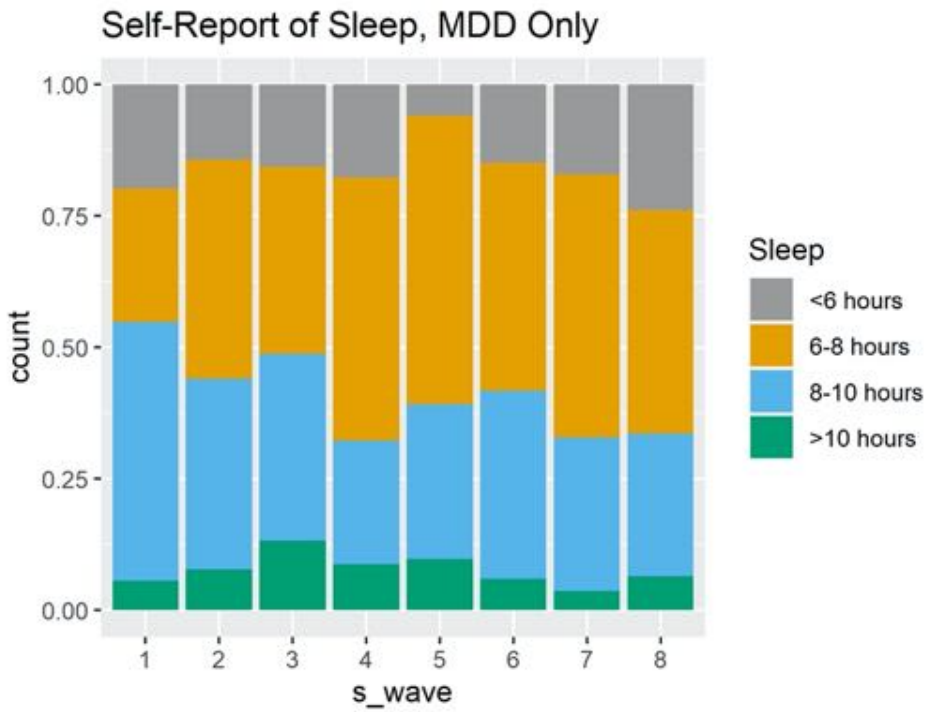




Preliminary results

Sleep pattern during COVID-19





Preliminary results

Received: 20 November 2017 | Revised: 1 June 2018 | Accepted: 14 June 2018
DOI: 10.1002/hbm.24278

WILEY

RESEARCH ARTICLE

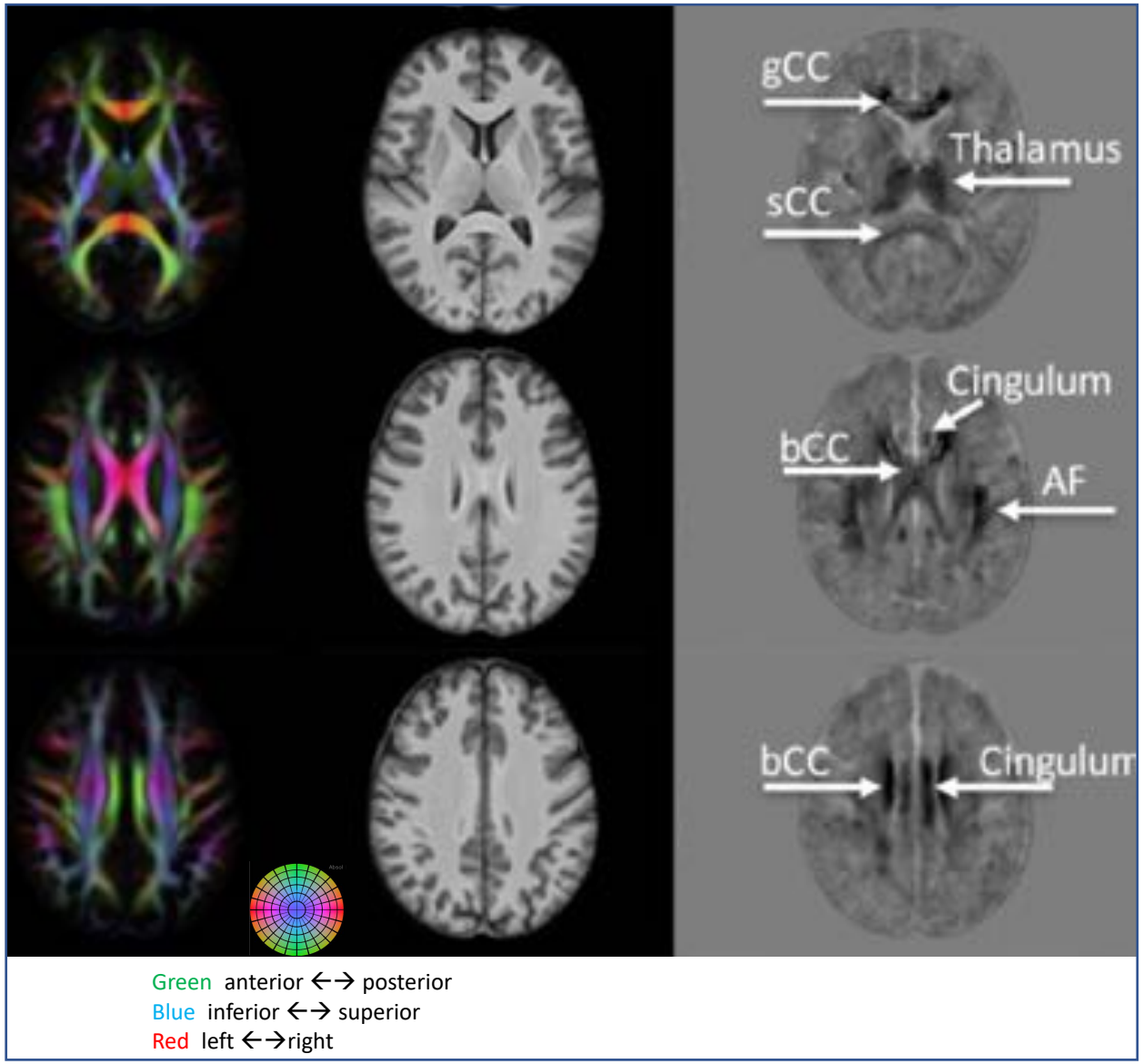
Tensor-based morphometry using scalar and directional information of diffusion tensor MRI data (DTBM): Application to hereditary spastic paraplegia

Neda Sadeghi[1] | Filippo Arrigoni[2] | Maria Grazia D'Angelo[3] | Cibu Thomas[4] | M. Okan Irfanoglu[1] | Elizabeth B. Hutchinson[1,5] | Amritha Nayak[1,5] | Pooja Modi[6] | Maria Teresa Bassi[7] | Carlo Pierpaoli[1]

Data

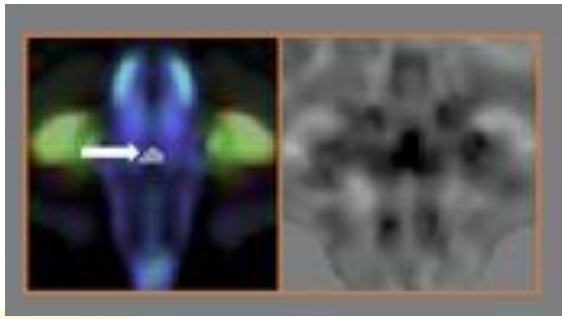
Twenty-four healthy volunteers (mean age of 35 and standard deviation of six years; 15 female and 9 male)

Four subjects diagnosed with Spastic paraplegia of type 11 (mutation in SPG11 gene) (mean age of 32 and standard deviation of three years; four female)



Control templates and effect size of LogJ maps are shown. Dark voxels in the effect size images represent regions where structures in the patient group are smaller than the control group (regions of atrophy or hypoplasia), whereas the bright regions, such as CSF spaces and ventricles, are areas that are larger in the patient group compared to the control group.

Is this region of reduced volume, an imaging marker shared among all MBS subjects?



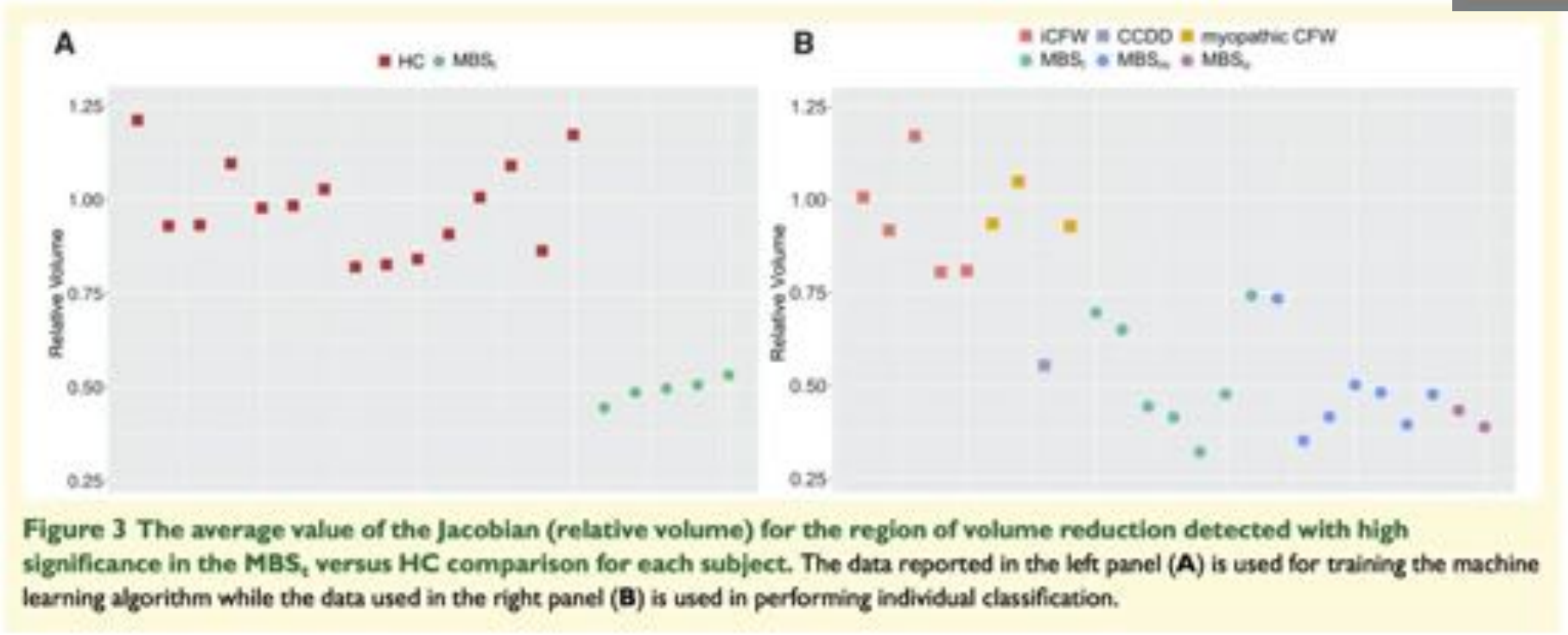


Figure 3 The average value of the Jacobian (relative volume) for the region of volume reduction detected with high significance in the MBS$_t$ versus HC comparison for each subject. The data reported in the left panel (A) is used for training the machine learning algorithm while the data used in the right panel (B) is used in performing individual classification.

Sadeghi et al., Brain Communications 2020

References

- ER Tufte (1983) The visual display of quantitative information. Graphics Press.
- https://rafalab.github.io/dsbook/data-visualization-principles.html
- https://r4ds.had.co.nz/introduction.html
- https://ggplot2.tidyverse.org
- https://www.rstudio.com/resources/cheatsheets/
- Interactive graphics:
  - https://shiny.rstudio.com/
  - https://d3js.org/

Section on Clinical and Computational Psychiatry

Thank you