Programming Solutions

John M. Drake & Ana I. Bento

Introduction

This document provides solutions to the Exercise "Programming".

Exercise. Write a script to load the West Nile virus data and use ggplot to create a histogram for the total number of cases in each state in each year. Follow the format of the *prototypical script* advocated in the presentation: Header, Load Packages, Declare Functions, Load Data, Perform Analysis.

```
library(tidyverse)
```

```
## Loading tidyverse: ggplot2
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: readr
## Loading tidyverse: purrr
## Loading tidyverse: dplyr
```

```
## Warning: package 'tibble' was built under R version 3.4.3
```

```
## Warning: package 'tidyr' was built under R version 3.4.3
```

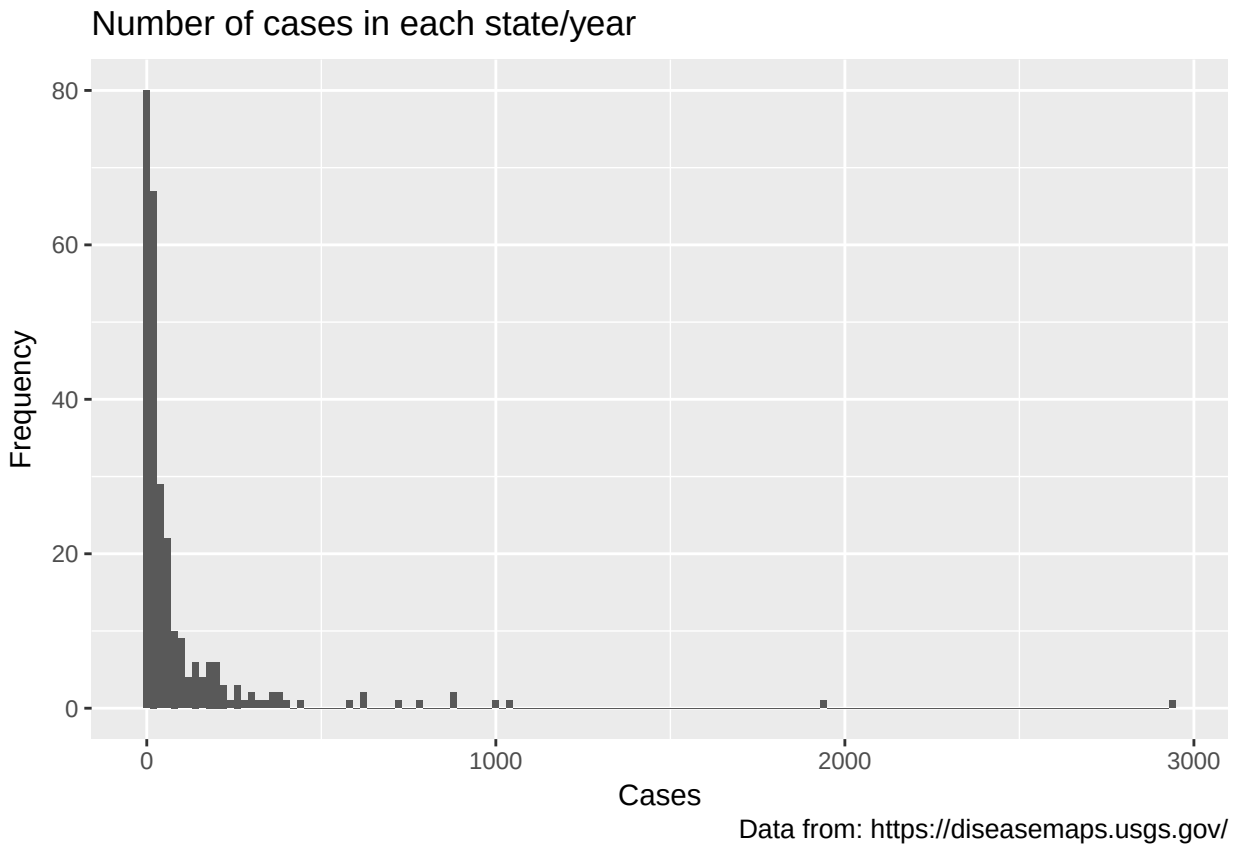
```
## Warning: package 'purrr' was built under R version 3.4.2
```

```
## Warning: package 'dplyr' was built under R version 3.4.2
```

```
## Conflicts with tidy packages ----------------------------------------------
```

```
## filter(): dplyr, stats
## lag():    dplyr, stats
```

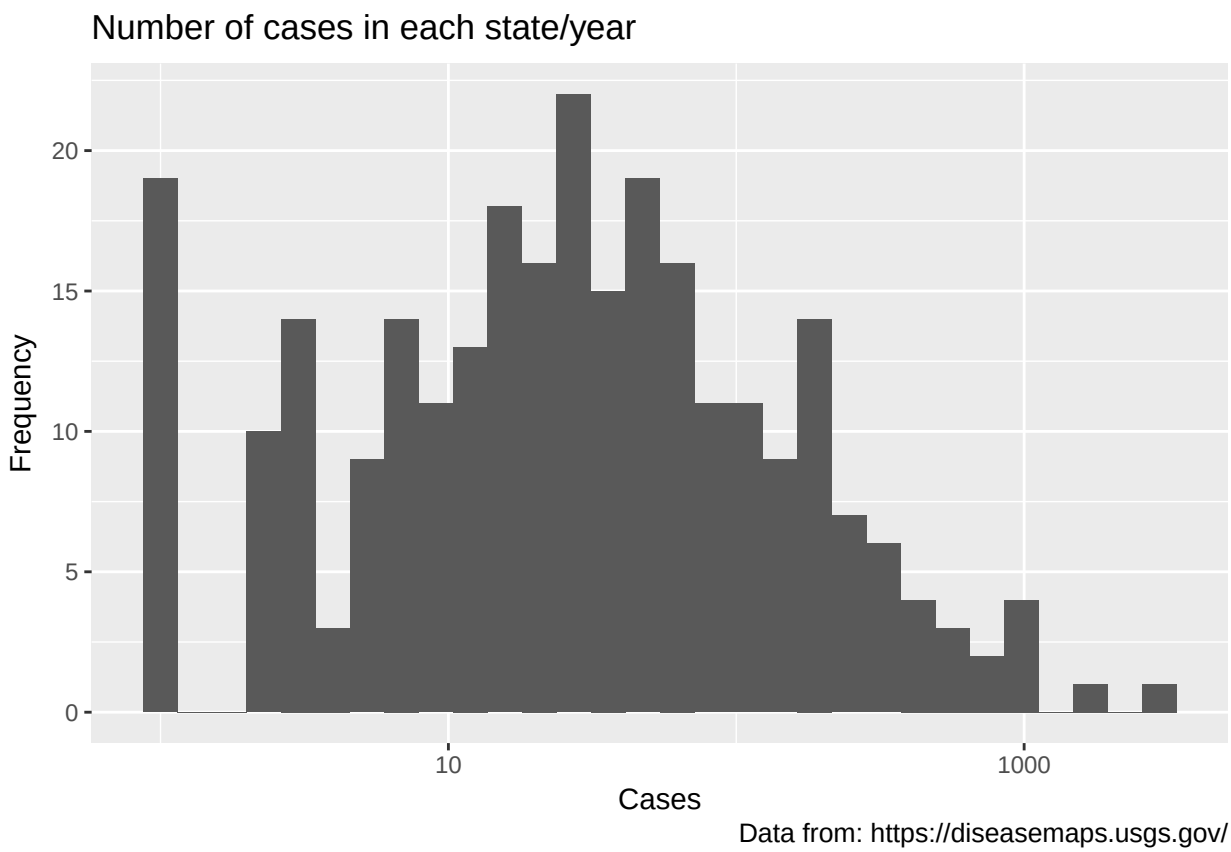
```
wnv <- read.csv('wnv.csv')
ggplot(data=wnv) +
  geom_histogram(mapping=aes(x=Total), binwidth = 20) +
  labs(x='Cases', y='Frequency',
       title='Number of cases in each state/year', caption="Data from: https://diseasemaps.usgs.gov/")
```



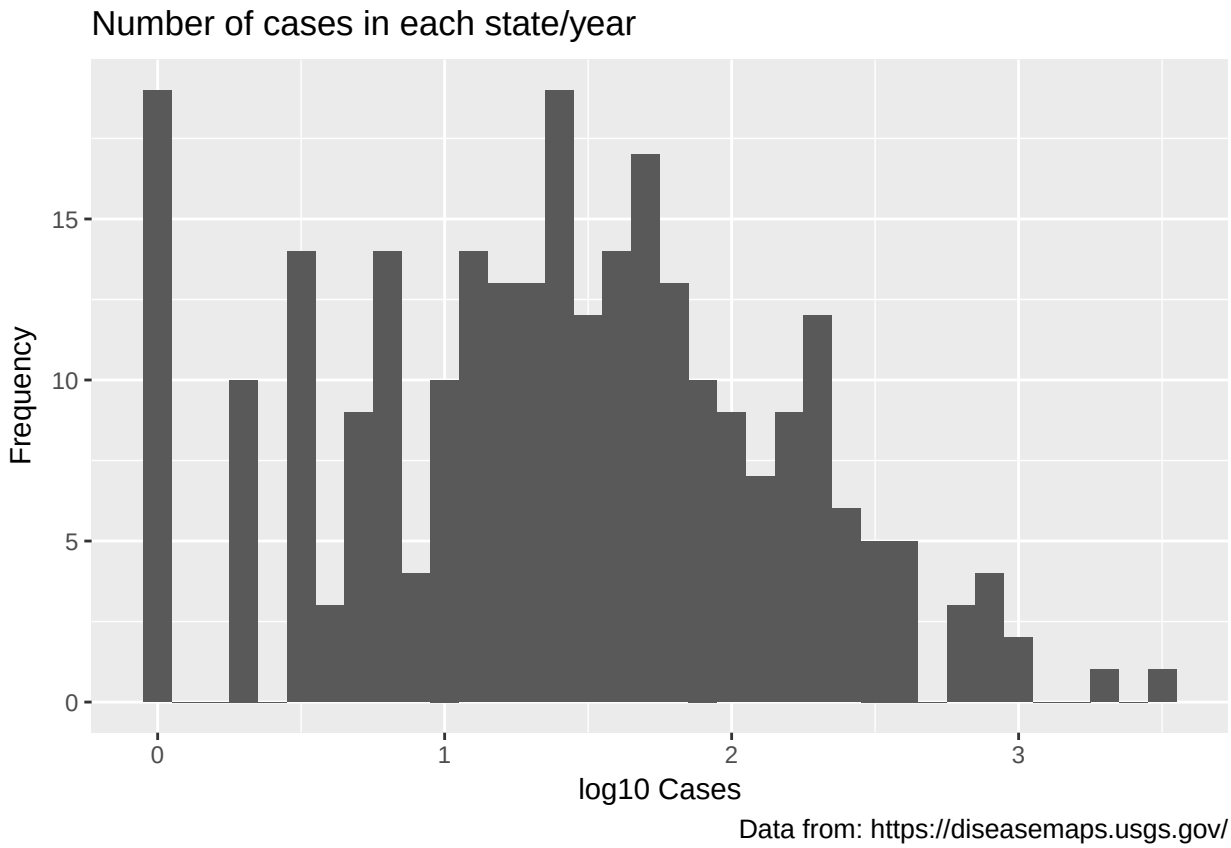
Exercise. The state-level and case burden is evidently highly skewed. Plot a histogram for the logarithm of the number of cases. Do this two different ways.

```
ggplot(data=wnv) +
  geom_histogram(mapping=aes(x=Total)) +
  scale_x_log10() +
  labs(x='Cases', y='Frequency',
       title='Number of cases in each state/year', caption="Data from: https://diseasemaps.usgs.gov/")
```

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

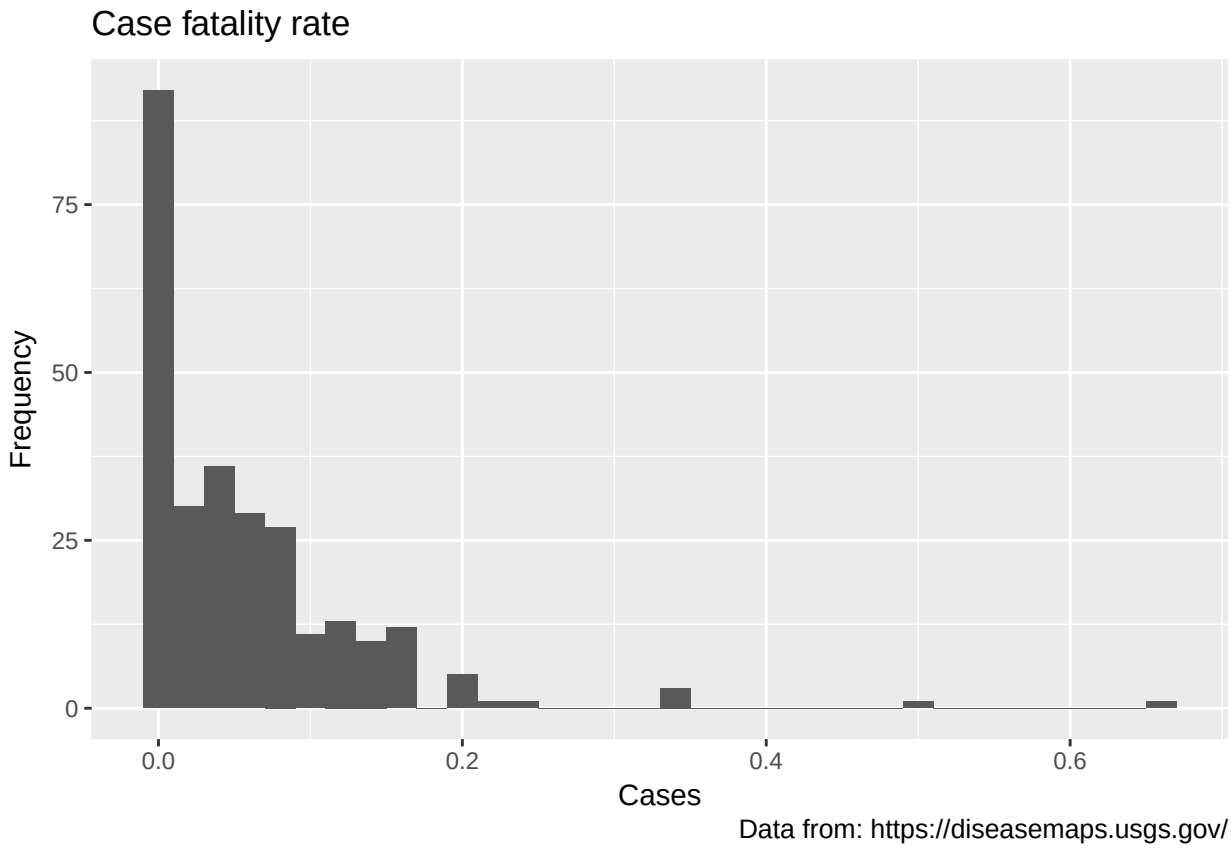



```
ggplot(data=wnv) +
  geom_histogram(mapping=aes(x=log(Total,10)), binwidth=0.1) +
  labs(x='log10 Cases', y='Frequency',
       title='Number of cases in each state/year', caption="Data from: https://diseasemaps.usgs.gov/")
```



Exercise. Use arithmetic operators to calculate the raw case fatality rate (CFR) in each state in each year. Plot a histogram of the calcated CFRs.

```
wnv$cfr <- wnv$Fatal / wnv$Total
ggplot(data=wnv) +
  geom_histogram(mapping=aes(x=cfr), binwidth=0.02) +
  labs(x='Cases', y='Frequency',
       title='Case fatality rate', caption="Data from: https://diseasemaps.usgs.gov/")
```



Exercise. **Use arithmetic operators, logical operators, and the function `sum` to verify that the variable `Total` is simply the sum of the number of febrile cases, neuroinvasive cases, and other cases.**

```
calculated.total <- wnv$Fever + wnv$EncephMen + wnv$Other
total.discrepancies <- sum(calculated.total != wnv$Total)
print(paste('Total discrepancies:', total.discrepancies))
```

```
## [1] "Total discrepancies: 0"
```

Exercise. **Use modular arithmetic to provide an annual case count for each state rounded (down) to the nearest dozen. Use modular arithmetic to extract the rounding errors associated with this calculate, then add the errors to obtain the total error.**

```
wnv$dozens <- wnv$Total %/% 12
wnv$errors <- wnv$Total %% 12
total.error <- sum(wnv$errors)
print(paste('Total error:', total.error))
```

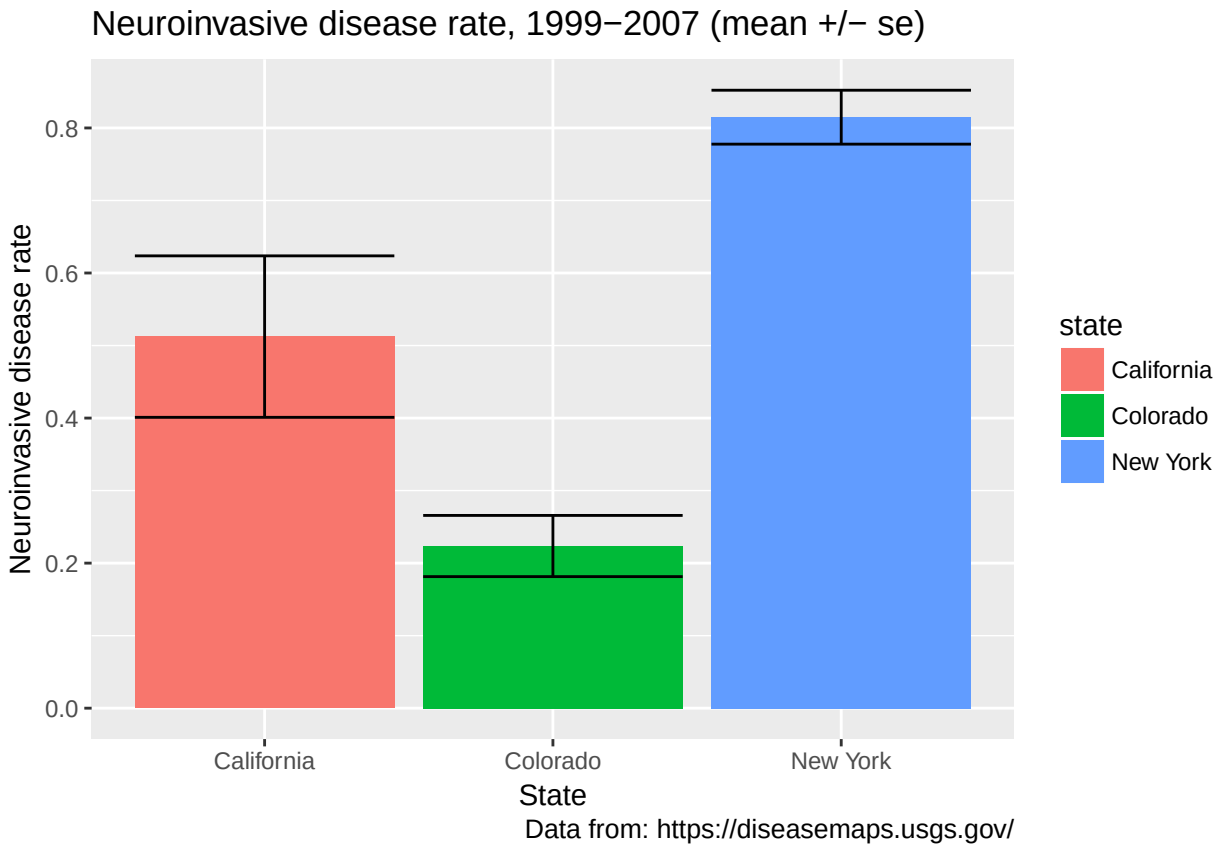
```
## [1] "Total error: 1241"
```

Exercise. **Let us call the ratio of meningitis/encephalitis cases to the total number of cases the *neuroinvasive disease rate*. Write a function to calculate the mean and standard error (standard deviation divided by the square root of the sample size) of the neuroinvasive disease rate for all the states in a given list and given set of years. Remember to place the function near the top of your script. Use your function to calculate the average severe disease rate in California, Colorado, and New York.**

Exercise. **Use ggplot to show the neurovinvasive disease rate for these states as a bar graph with error bars to show the standard deviation.**

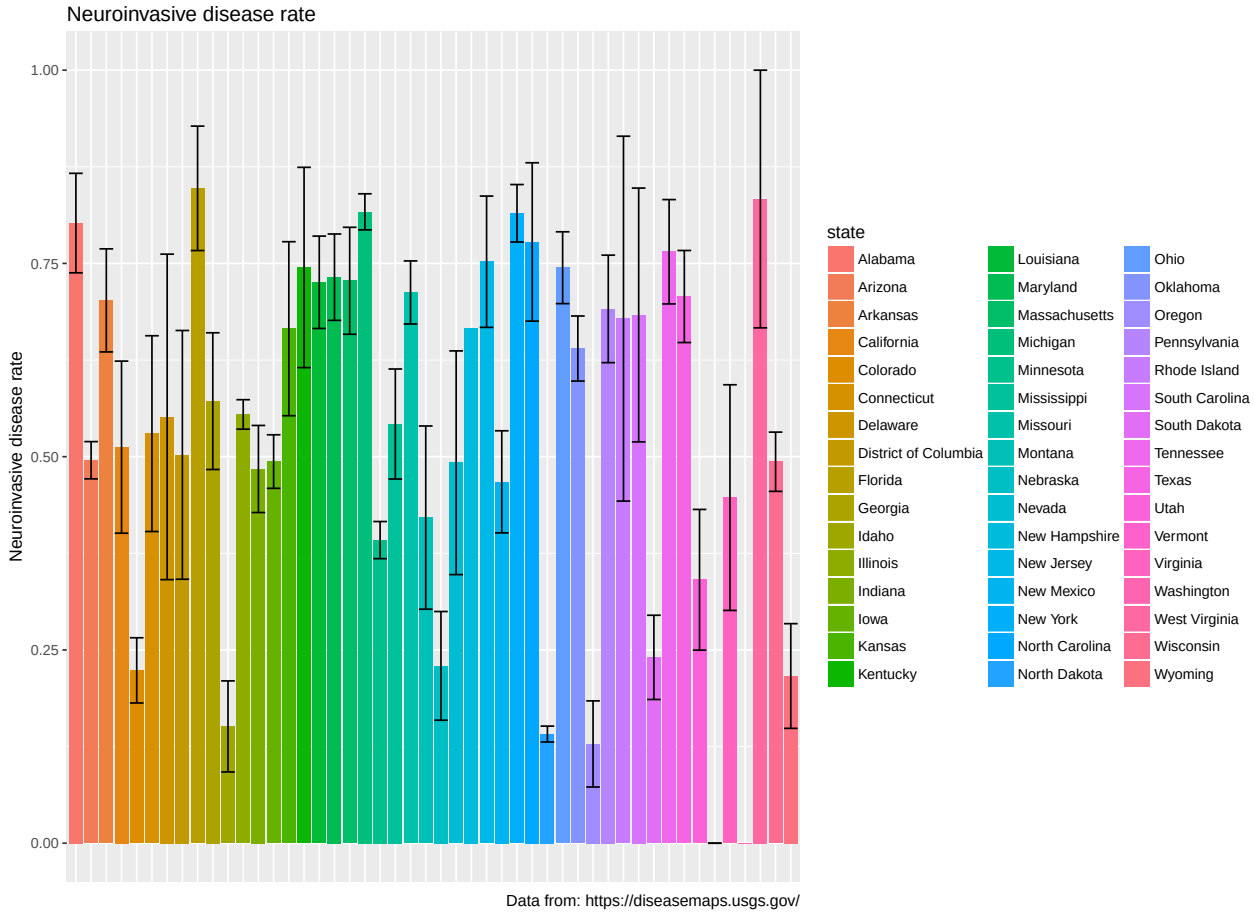
Exercise. Use your function and ggplot to show the neurovinvasive disease rate for all states.

```
ndr <- function(state='Colorado', years=1999:2007){
  # Computes mean and standard error of neuroinvasive disease rate by state
  #
  # Args:
  #  state: vector of state names
  #  years: vector of years to be included in the calculation
  #
  # Returns:
  #  a dataframe containing state name, mean neuroinvasive disease rate, and se

  x <- wnv[wnv$State %in% state & wnv$Year %in% years,]
  y <- data.frame(state = x$State, ndr = x$EncephMen / x$Total)
  m <- aggregate(y$ndr, by=list(y$state), FUN = mean)
  se <- aggregate(y$ndr, by=list(y$state), FUN = function(x) sd(x)/sqrt(length(x)) )
  out <- merge(m, se, by = 'Group.1')
  names(out) <- c('state','mean.ndr','se.ndr')
  return(out)
}

disease <- ndr(state=c('California','Colorado','New York'))

ggplot(disease, aes(x=state, y=mean.ndr, fill=state)) +
    geom_bar(stat="identity") +
    geom_errorbar(aes(ymin=mean.ndr-se.ndr, ymax=mean.ndr+se.ndr)) +
    labs(x='State', y='Neuroinvasive disease rate',
       title='Neuroinvasive disease rate, 1999-2007 (mean +/- se)', caption="Data from: https://diseasen
```




```
disease <- ndr(state=unique(wnv$State))

ggplot(disease, aes(x=state, y=mean.ndr, fill=state)) +
  geom_bar(stat="identity") +
  geom_errorbar(aes(ymin=mean.ndr-se.ndr, ymax=mean.ndr+se.ndr)) +
  labs(x='State', y='Neuroinvasive disease rate',
     title='Neuroinvasive disease rate', caption="Data from: https://diseasemaps.usgs.gov/") +
  theme(axis.title.x=element_blank(),
        axis.text.x=element_blank(),
        axis.ticks.x=element_blank())
```

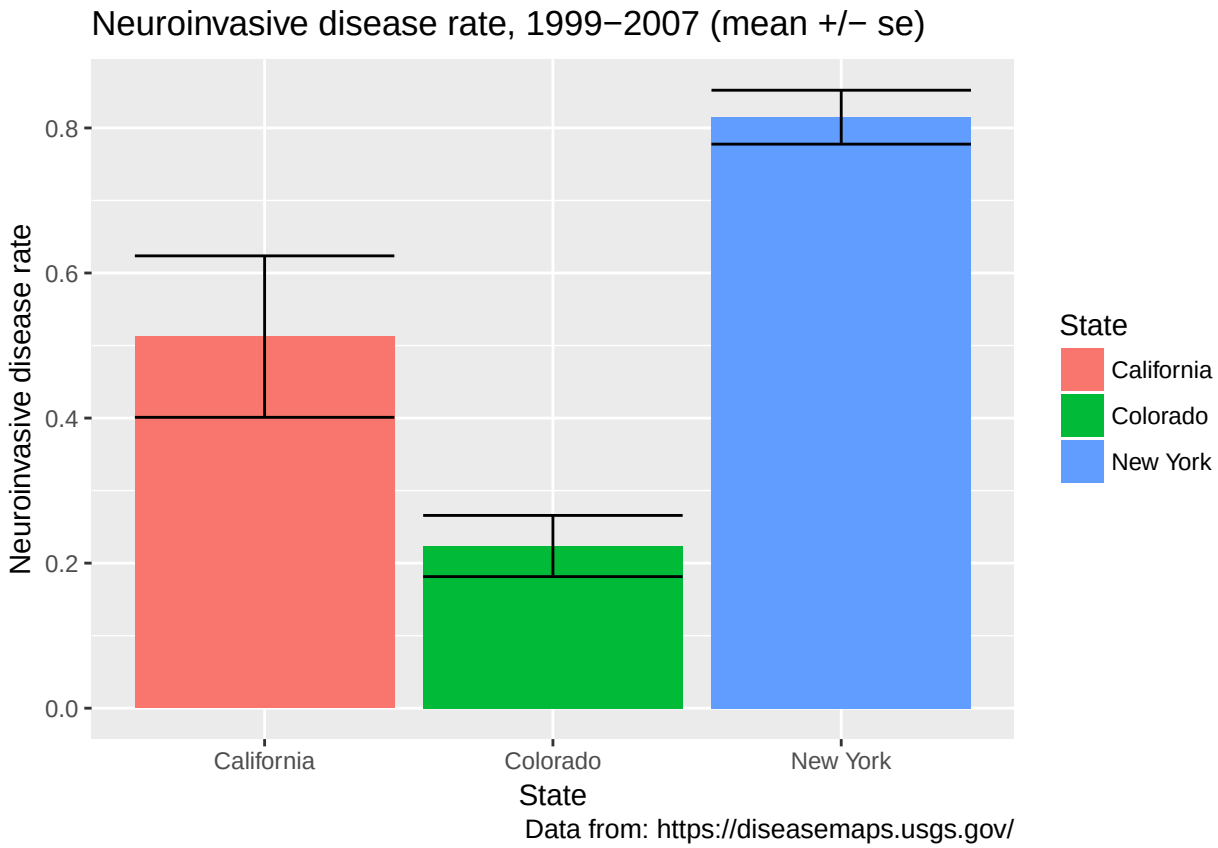
```
## Warning: Removed 2 rows containing missing values (geom_errorbar).
```



Exercise. Use pipes to produce the same plots without using your function.

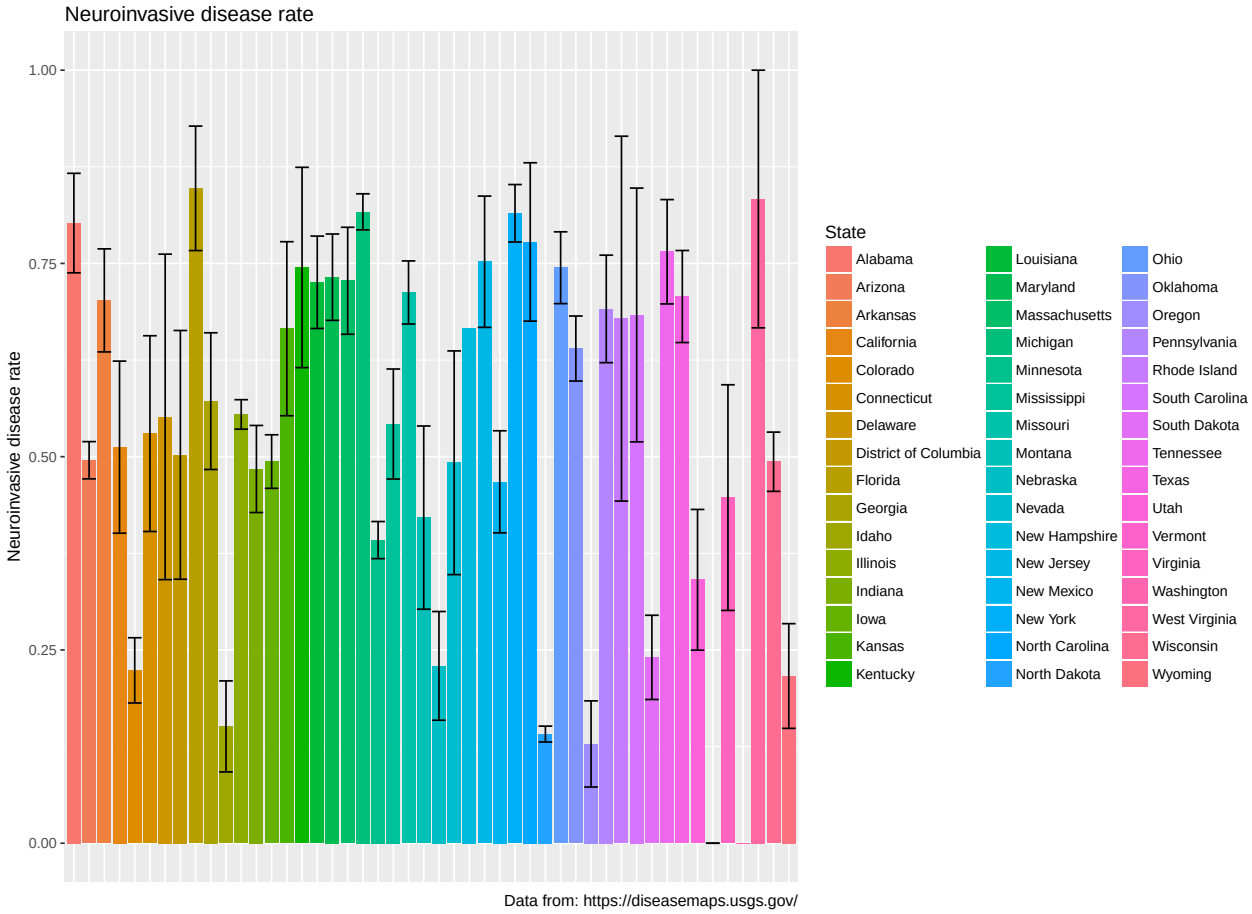
```
library(dplyr)
wnv %>%
  filter(State %in% c('California','Colorado','New York')) %>%
  group_by(State) %>%
  summarize(mean.ndr = mean(EncephMen/Total), se.ndr = sd(EncephMen/Total)/sqrt(length(EncephMen/Total)))
  ggplot(aes(x=State, y=mean.ndr, fill=State)) +
  geom_bar(stat="identity") +
  geom_errorbar(aes(ymin=mean.ndr-se.ndr, ymax=mean.ndr+se.ndr)) +
  labs(x='State', y='Neuroinvasive disease rate',
     title='Neuroinvasive disease rate, 1999-2007 (mean +/- se)', caption="Data from: https://diseasemap
```

```
## Warning: package 'bindrcpp' was built under R version 3.4.4
```




```
wnv %>%
  filter(State %in% unique(State)) %>%
  group_by(State) %>%
  summarize(mean.ndr = mean(EncephMen/Total), se.ndr = sd(EncephMen/Total)/sqrt(length(EncephMen/Total)))
  ggplot(aes(x=State, y=mean.ndr, fill=State)) +
  geom_bar(stat="identity") +
  geom_errorbar(aes(ymin=mean.ndr-se.ndr, ymax=mean.ndr+se.ndr)) +
  labs(x='State', y='Neuroinvasive disease rate',
     title='Neuroinvasive disease rate', caption="Data from: https://diseasemaps.usgs.gov/") +
  theme(axis.title.x=element_blank(),
        axis.text.x=element_blank(),
        axis.ticks.x=element_blank())
```

```
## Warning: Removed 2 rows containing missing values (geom_errorbar).
```



To be continued...