

STAT540 Homework 5 Solution

Question 1

The problem with the plot is best described by option b: The axis does not start at 0. Because the vertical axis starts at 200 instead of 0, the visual difference between the bars is exaggerated.

Question 2

The correct answer is c: The plot on the right is better because alphabetical order has nothing to do with the disease, and ordering according to the actual rate quickly shows the states with the highest and lowest rates. This arrangement allows for a more straightforward comparison, making it easy to identify the states with the highest and lowest measles rates at a glance.

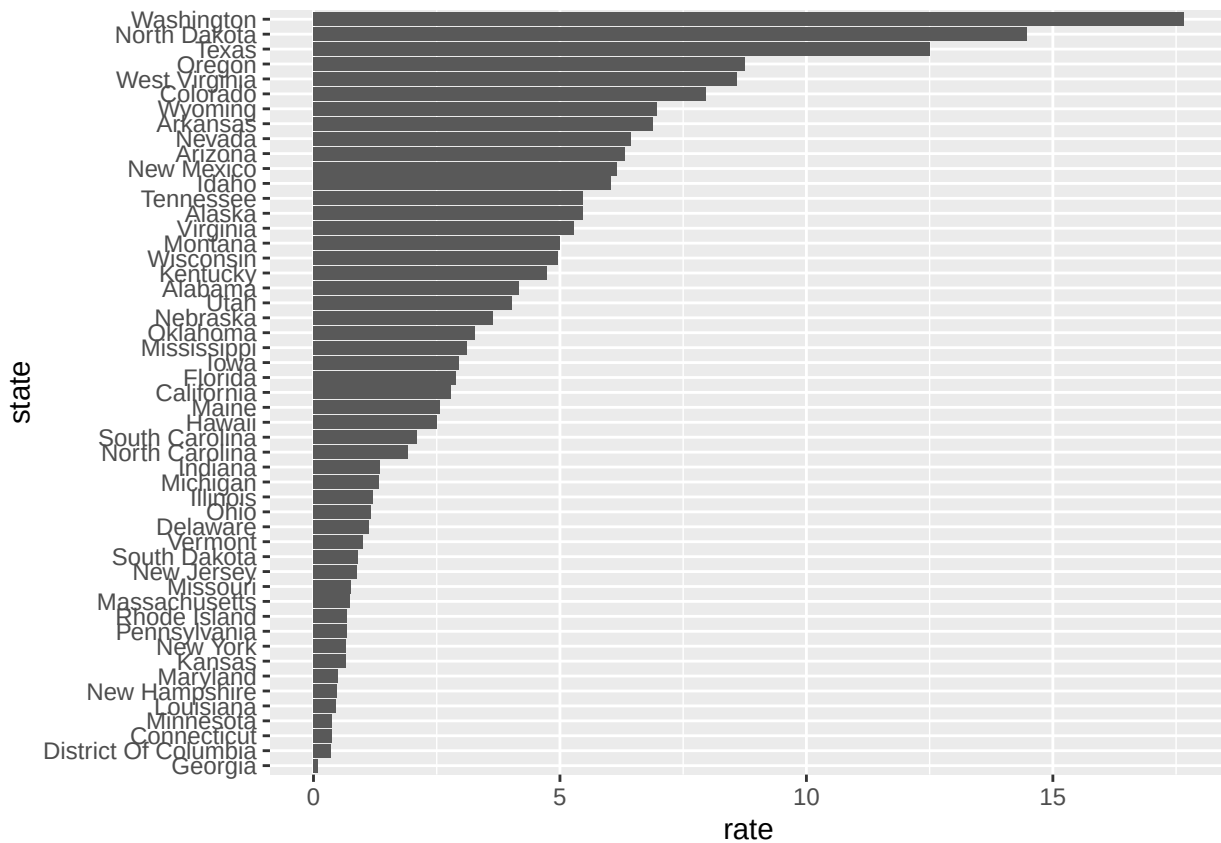
Question 3

```
library(dslabs)
library(dplyr)

## Warning: package 'dplyr' was built under R version 4.4.3
##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:stats':
##
##   filter, lag
## The following objects are masked from 'package:base':
##
##   intersect, setdiff, setequal, union
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.4.3
dat <- us_contagious_diseases %>%
  filter(year == 1967 & disease == "Measles" & !is.na(population)) %>%
  mutate(rate = count / population * 10000 * 52 / weeks_reporting)

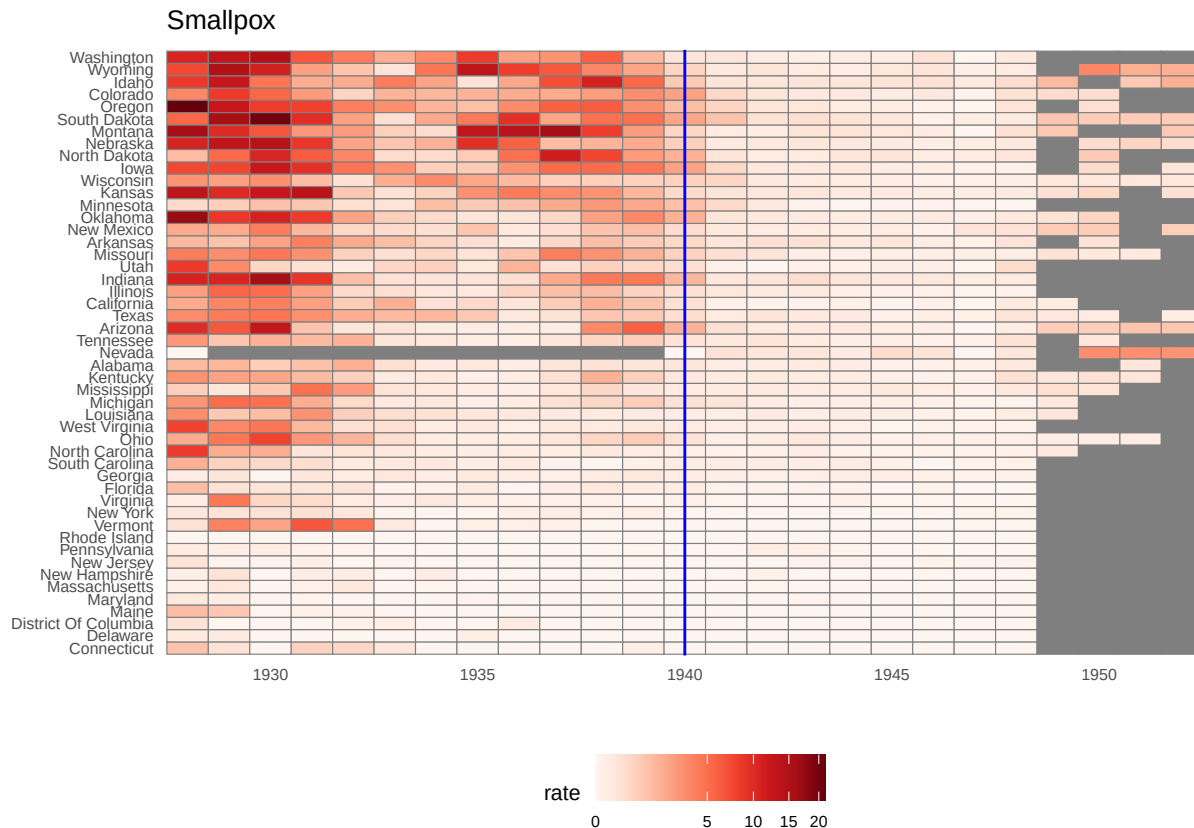
dat |>
  mutate(state = forcats::fct_reorder(state, rate, .desc = F)) |>
  ggplot(aes(state, rate)) +
  geom_col() +
  coord_flip()
```



Question 4

```
library(RColorBrewer)
the_disease <- "Smallpox"
dat <- us_contagious_diseases |>
  filter(!state %in% c("Hawaii", "Alaska") & disease == the_disease) |>
  mutate(rate = count / population * 10000 * 52 / weeks_reporting) |>
  mutate(state = reorder(state, ifelse(year <= 1963, rate, NA),
    median, na.rm = TRUE))

dat |> ggplot(aes(year, state, fill = rate)) +
  geom_tile(color = "grey50") +
  scale_x_continuous(expand = c(0,0)) +
  scale_fill_gradientn(colors = brewer.pal(9, "Reds"), trans = "sqrt") +
  geom_vline(xintercept = 1940, col = "blue") +
  theme_minimal() +
  theme(panel.grid = element_blank(),
    legend.position = "bottom",
    text = element_text(size = 8)) +
  labs(title = the_disease, x = "", y = "")
```



Question 5

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
## v forcats 1.0.0      v stringr 1.5.1
## v lubridate 1.9.3    v tibble 3.2.1
## v purrr 1.0.2       v tidyr 1.3.1
## v readr 2.1.5
```

```
## -- Conflicts ----- tidyverse_conflicts() --
```

```
## x dplyr::filter() masks stats::filter()
```

```
## x dplyr::lag() masks stats::lag()
```

```
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(dslabs)
```

```
library(RColorBrewer)
```

```
data("us_contagious_diseases")
```

```
#the_disease <- "Measles"
```

```
dat <- us_contagious_diseases |>
```

```
  filter(state %in% c("California") & weeks_reporting >= 10) |>
```

```
  mutate(rate = count / population * 10000 * 52 / weeks_reporting) |>
```

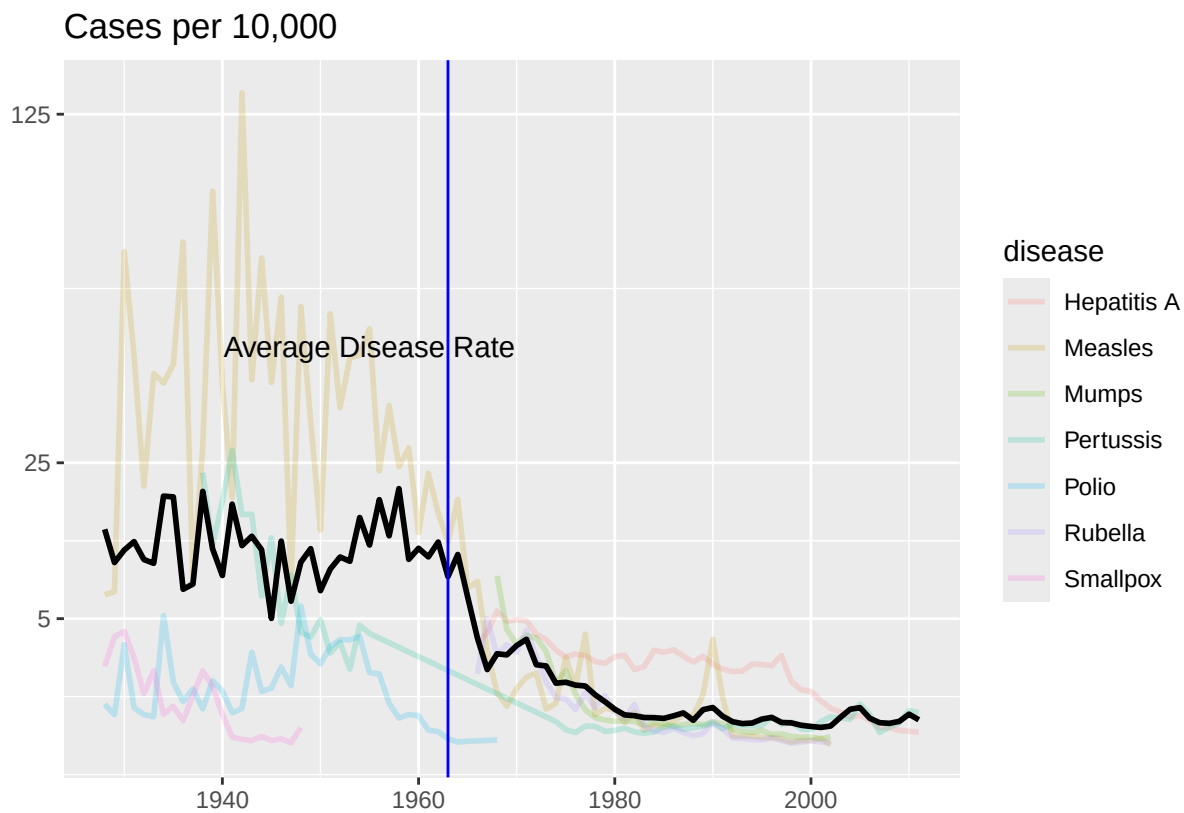
```
  mutate(state = reorder(state, ifelse(year <= 1963, rate, NA),
                               median, na.rm = TRUE))
```

```

avg <- us_contagious_diseases |> group_by(year) |>
  summarize(us_rate = sum(count, na.rm = TRUE) /
    sum(population, na.rm = TRUE) * 10000)

dat |>
  filter(!is.na(rate)) |>
  ggplot() +
  geom_line(aes(year, rate, group = disease, color = disease),
    show.legend = T, alpha = 0.2, linewidth = 1) +
  geom_line(mapping = aes(year, us_rate), data = avg, linewidth = 1) +
  scale_y_continuous(trans = "sqrt", breaks = c(5, 25, 125, 300)) +
  ggtitle("Cases per 10,000") +
  xlab("") + ylab("") +
  geom_text(data = data.frame(x = 1955, y = 50),
    mapping = aes(x, y, label = "Average Disease Rate"),
    color = "black") +
  geom_vline(xintercept = 1963, col = "blue")

```



Question 6

(i)

The class of p is gg" andggplot".

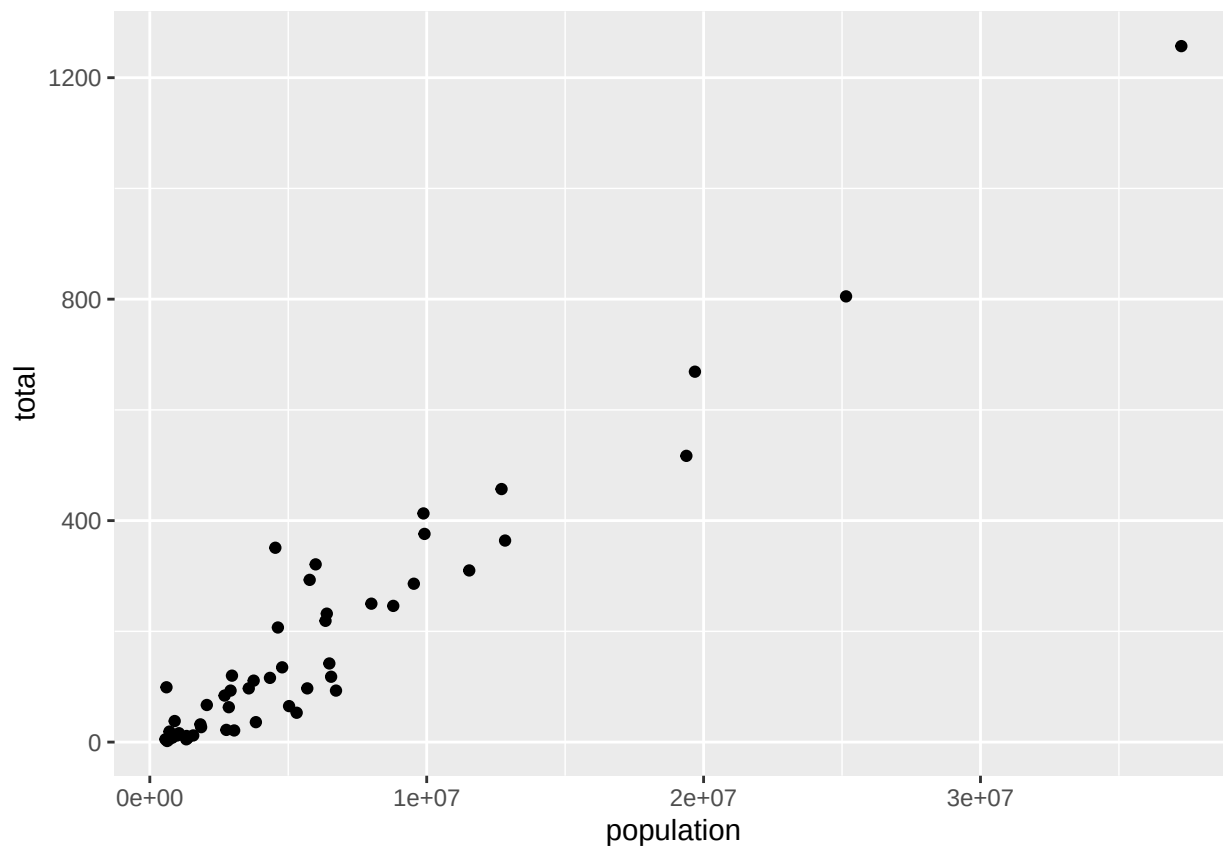
```
library(dplyr)
library(ggplot2)
library(dslabs)

p <- ggplot(data = murders)
class(p)

## [1] "ggplot2::ggplot" "ggplot"          "ggplot2::gg"     "S7_object"
## [5] "gg"
```

(ii)

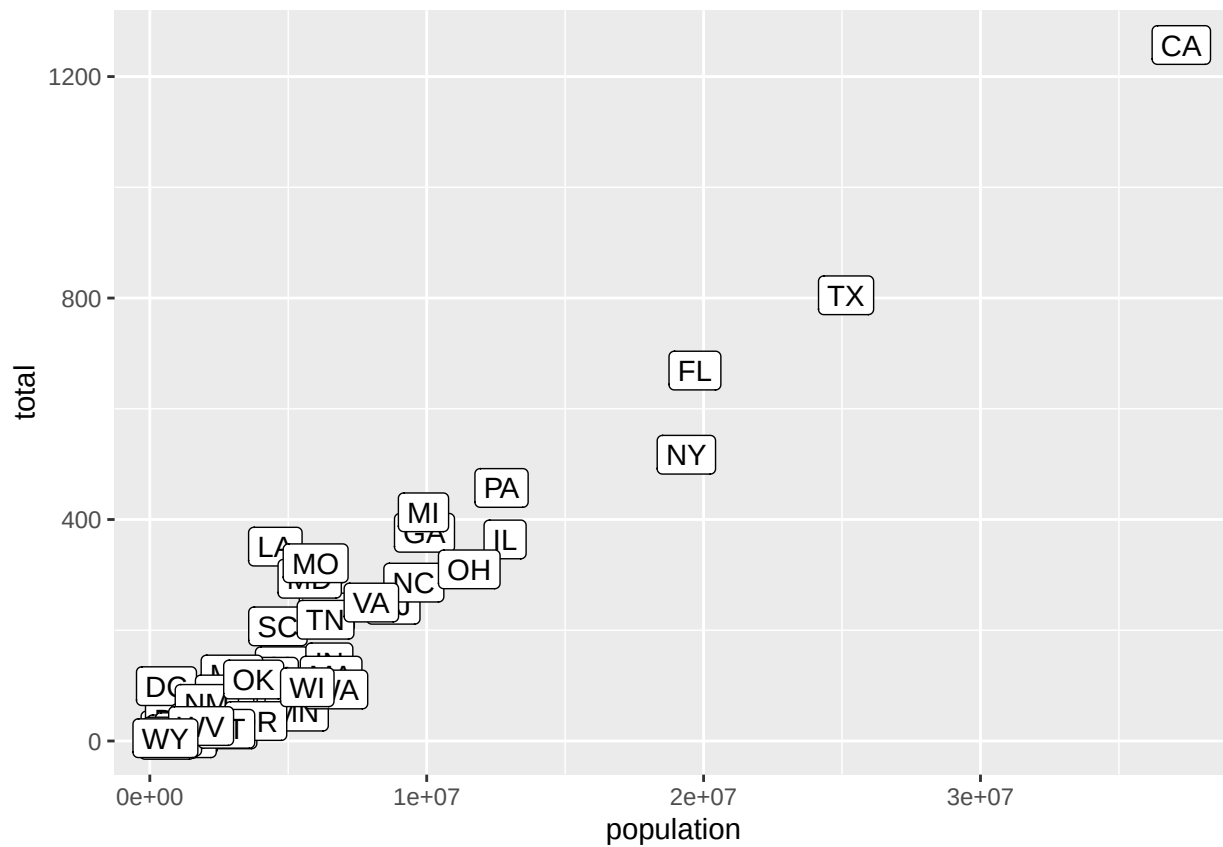
```
murders |> ggplot(aes(x = population, y = total)) +
  geom_point()
```



(iii)

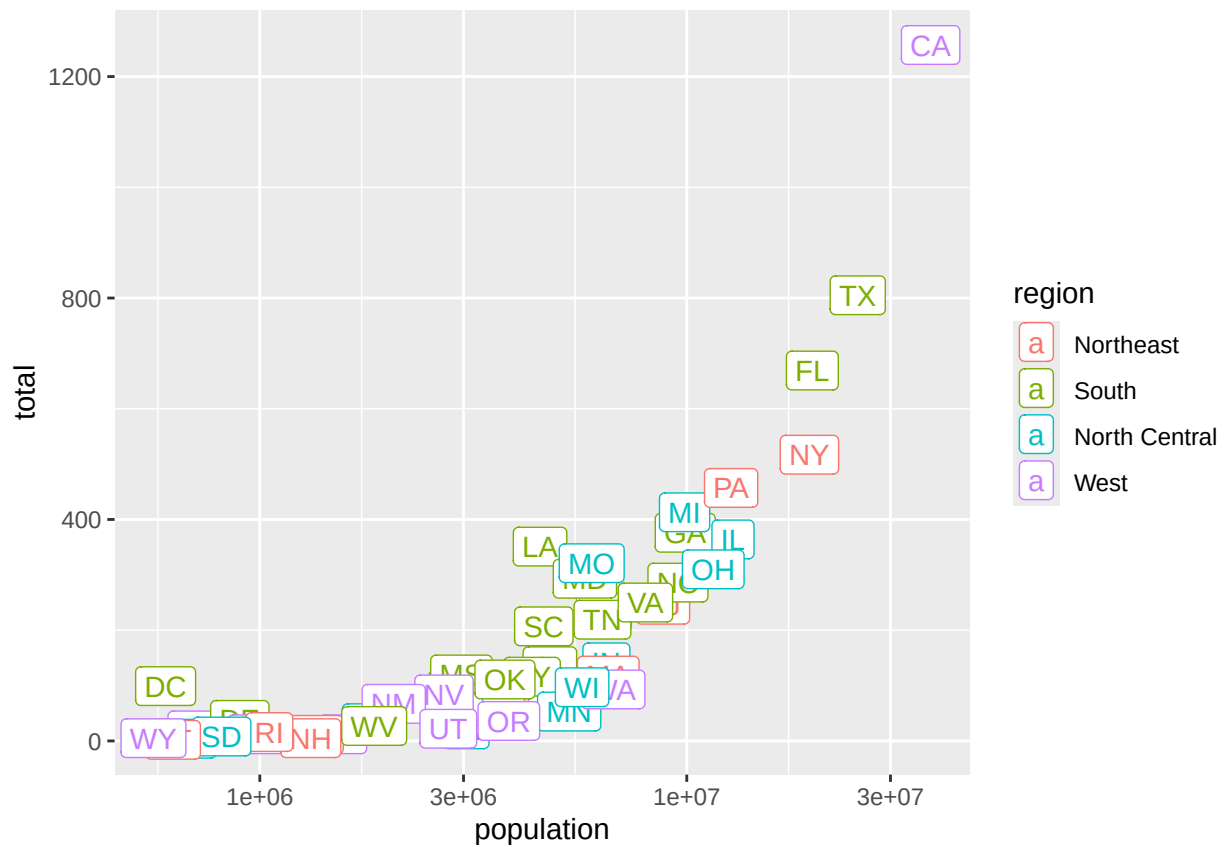
The answer is a.

```
murders |> ggplot(aes(population, total, label = abb)) +
  geom_label()
```



(iv)

```
p <- murders |> ggplot(aes(population, total, label = abb, color = region)) +  
  geom_label()  
p + scale_x_log10()
```



Question 7

```
heights |> ggplot(aes(height, fill = sex)) +
  geom_density(alpha = 0.2)
```

