Homework 2_Suixin Jiang

Suixin Jiang

09/26/2019

Exercise 1: Msleep data

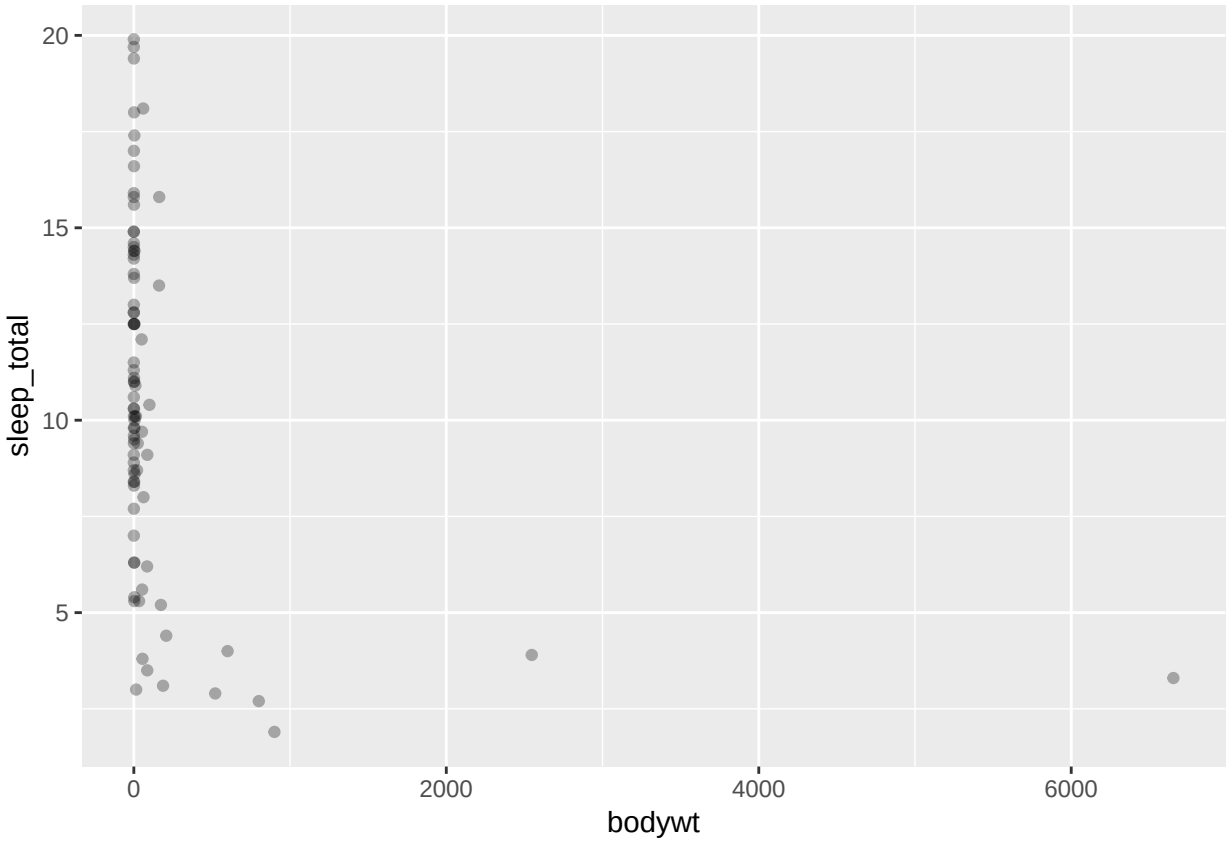
Q1 & Q2: There are 83 mammals and 11 variables in the msleep data frame.

```
data(msleep); str(msleep)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    83 obs. of  11 variables:
##  $ name        : chr  "Cheetah" "Owl monkey" "Mountain beaver" "Greater short-tailed shrew" ...
##  $ genus       : chr  "Acinonyx" "Aotus" "Aplodontia" "Blarina" ...
##  $ vore        : chr  "carni" "omni" "herbi" "omni" ...
##  $ order       : chr  "Carnivora" "Primates" "Rodentia" "Soricomorpha" ...
##  $ conservation: chr  "lc" NA "nt" "lc" ...
##  $ sleep_total : num  12.1 17 14.4 14.9 4 14.4 8.7 7 10.1 3 ...
##  $ sleep_rem   : num  NA 1.8 2.4 2.3 0.7 2.2 1.4 NA 2.9 NA ...
##  $ sleep_cycle : num  NA NA NA 0.133 0.667 ...
##  $ awake       : num  11.9 7 9.6 9.1 20 9.6 15.3 17 13.9 21 ...
##  $ brainwt     : num  NA 0.0155 NA 0.00029 0.423 NA NA NA 0.07 0.0982 ...
##  $ bodywt      : num  50 0.48 1.35 0.019 600 ...
```

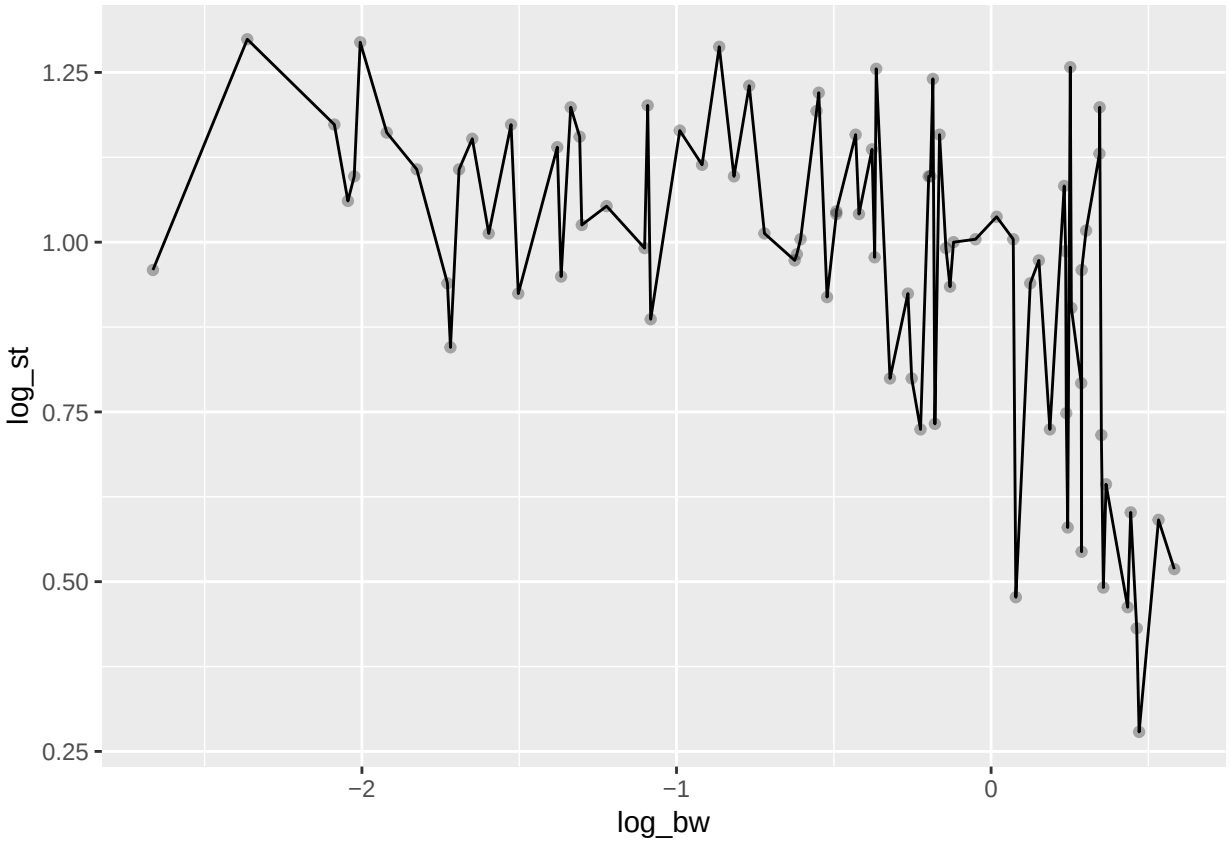
Q3: Since most mammals have a light-weight, the scatter plot is very condensed in the left.

```
ggplot(msleep, mapping = aes(x = bodywt, y = sleep_total)) +
  geom_point(alpha = 0.3, na.rm = T)
```



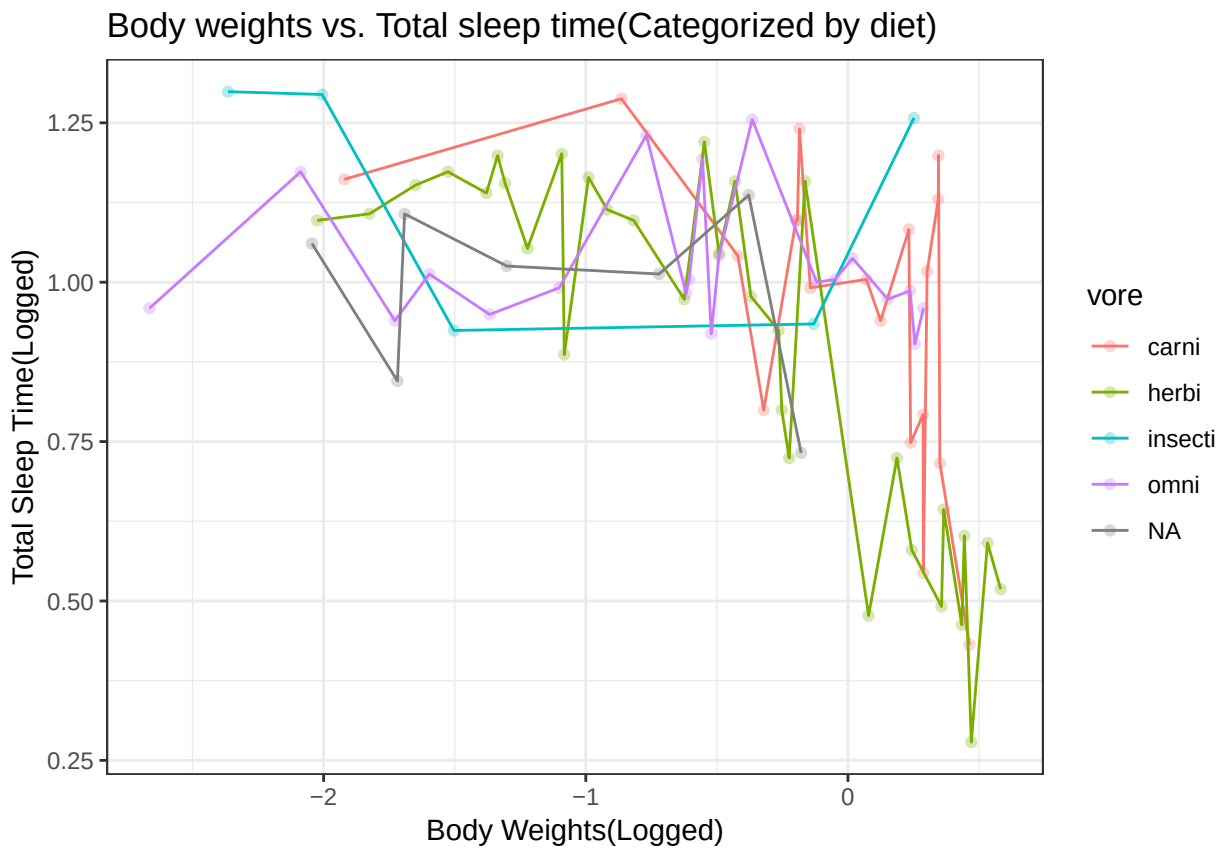
Q4: After a log transformation, the plot looks better.

```
log_st = log10(msleep$sleep_total)
log_bw = log10(log10(msleep$bodywt + 1))
ggplot(msleep, mapping = aes(x = log_bw, y = log_st)) +
  geom_point(alpha = 0.3, na.rm = T) +
  geom_line()
```



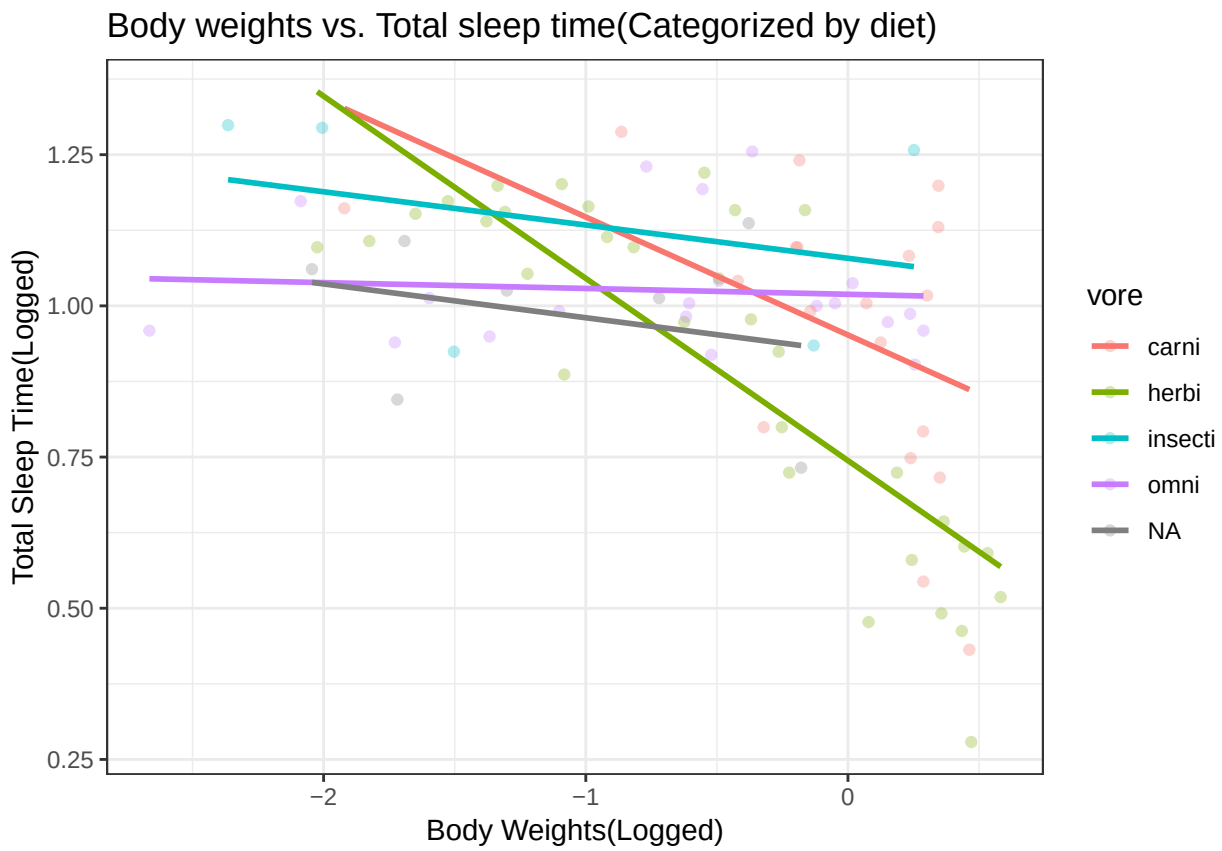
Q5: Q4 plot adds on axis labels, title, and theme.

```
ggplot(msleep, mapping = aes(x = log_bw, y = log_st, color = vore)) +
  geom_point(alpha = 0.3) +
  geom_line() +
  theme_bw() +
  xlab('Body Weights(Logged)') +
  ylab('Total Sleep Time(Logged)') +
  ggtitle('Body weights vs. Total sleep time(Categorized by diet)')
```



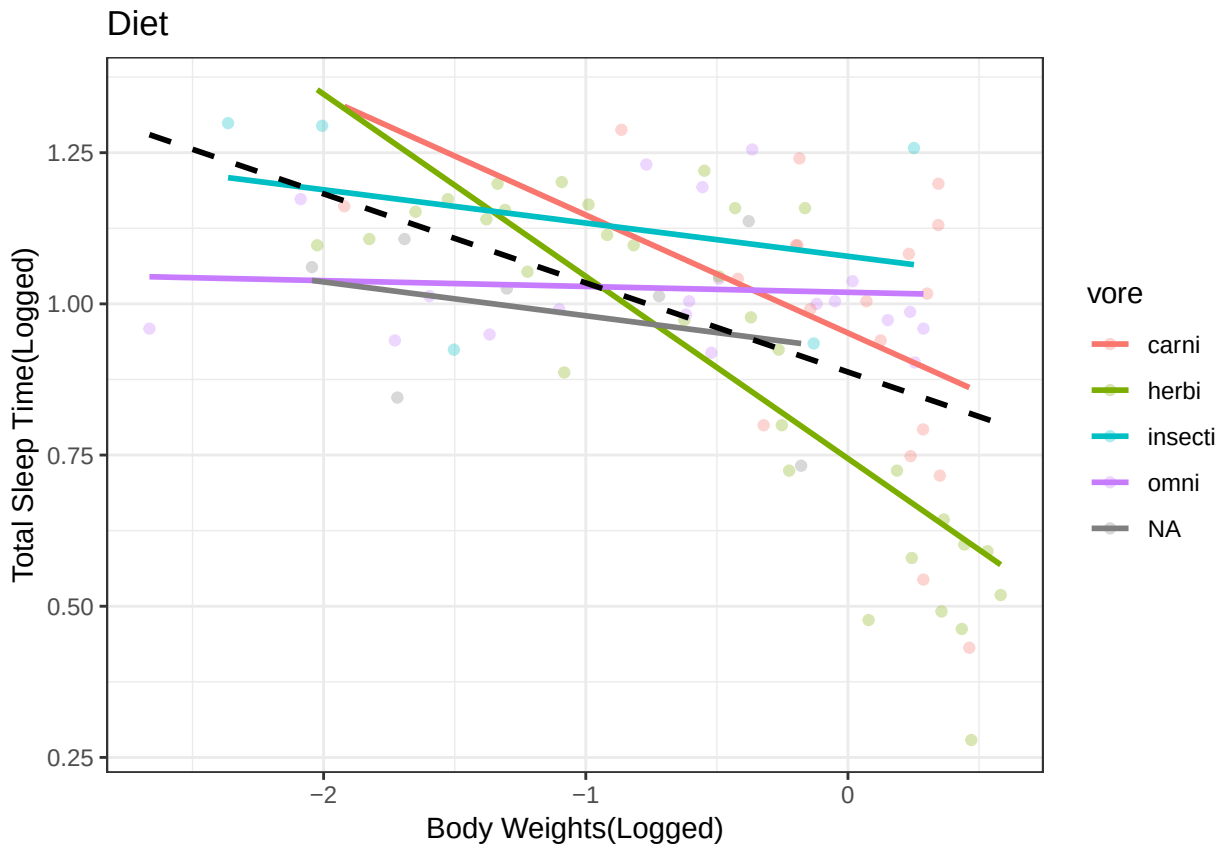
Q6: Herbivore and Carnivore body weights have a significant effect on total sleep time.

```
ggplot(msleep, mapping = aes(x = log_bw, y = log_st, color = vore)) +
  geom_point(alpha = 0.3) +
  theme_bw() +
  xlab('Body Weights(Logged)') +
  ylab('Total Sleep Time(Logged)') +
  ggtitle('Body weights vs. Total sleep time(Categorized by diet)') +
  geom_smooth(se = F, method = lm)
```



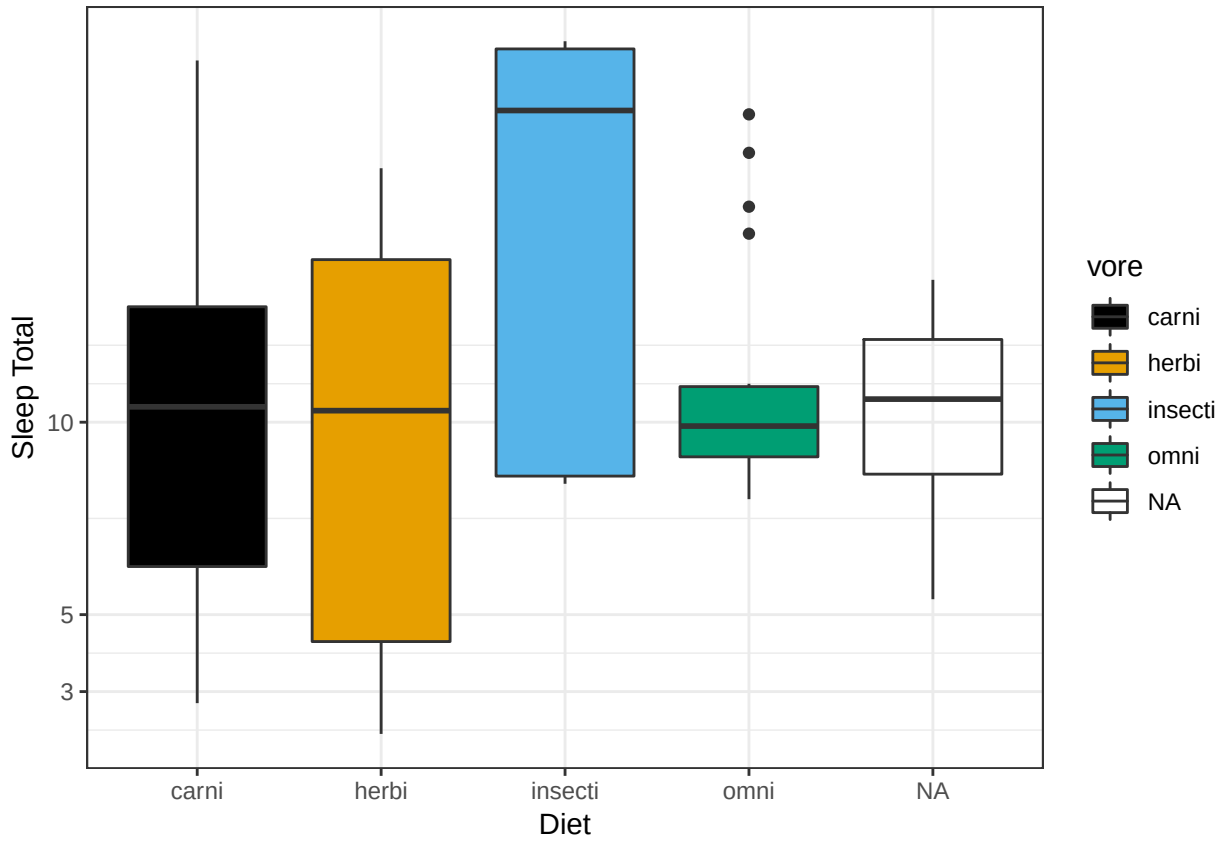
Q7 & Q8: Overall OLS dashed line adds on previous plot with 'Diet' title.

```
ggplot(msleep, mapping = aes(x = log_bw, y = log_st, color = vore)) +
  geom_point(alpha = 0.3) +
  theme_bw() +
  xlab('Body Weights(Logged)') +
  ylab('Total Sleep Time(Logged)') +
  ggtitle('Diet') +
  geom_smooth(se = F, method = lm) +
  geom_smooth(se = F, method = lm, linetype = 'dashed', color = 'black')
```

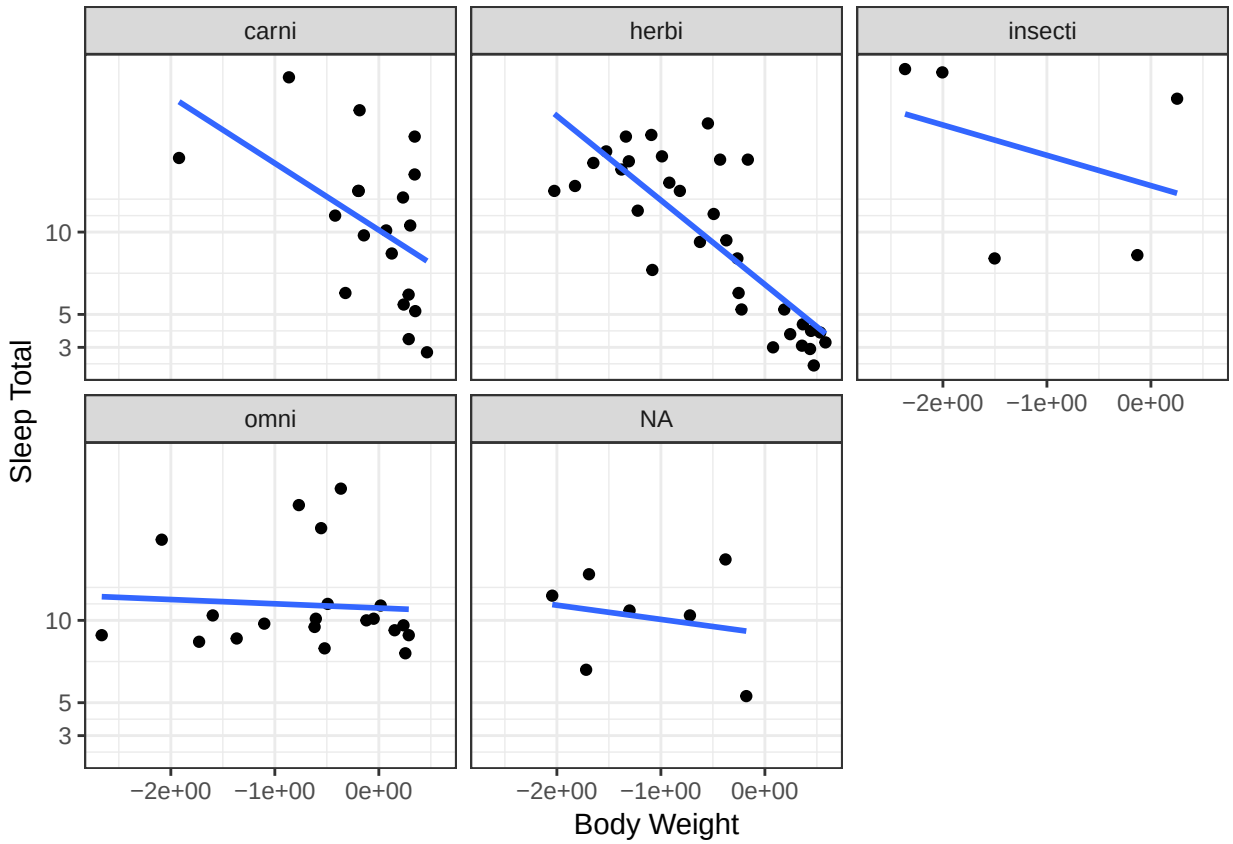


Q9: Reproduced plot. And due to different log transformation, the plots are not exactly the same.

```
ggplot(msleep, mapping = aes(x = vore, y = sleep_total, fill = vore)) +
  geom_boxplot() +
  theme_bw() +
  scale_fill_colorblind() +
  scale_y_continuous('Sleep Total', breaks = c(3,5,10)) +
  xlab('Diet')
```




```
ggplot(msleep, mapping = aes(x = log_bw, y = sleep_total)) +
  geom_point() +
  geom_smooth(se = F, method = lm) +
  theme_bw() +
  facet_wrap( ~ vore) +
  scale_x_continuous('Body Weight', labels = scientific) +
  scale_y_continuous('Sleep Total', breaks = c(3,5,10))
```



Exercise 2: Midwest Data

Q1 & Q2: The observational units of 'midwest' data frame are the number of population of each midwest county in the U.S. and some other feature ratios.

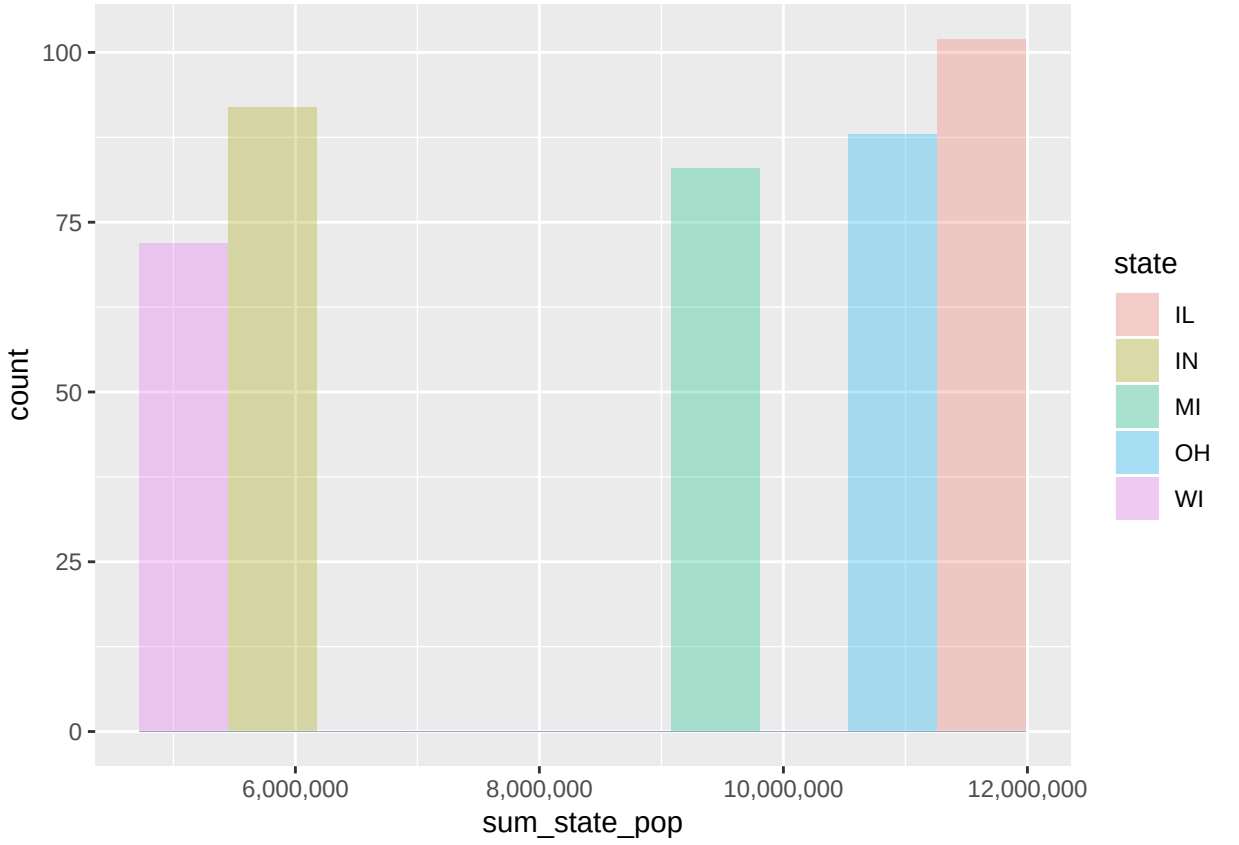
```
data(midwest); str(midwest)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    437 obs. of  28 variables:
##  $ PID                 : int  561 562 563 564 565 566 567 568 569 570 ...
##  $ county              : chr  "ADAMS" "ALEXANDER" "BOND" "BOONE" ...
##  $ state               : chr  "IL" "IL" "IL" "IL" ...
##  $ area                : num  0.052 0.014 0.022 0.017 0.018 0.05 0.017 0.027 0.024 0.058 ...
##  $ poptotal            : int  66090 10626 14991 30806 5836 35688 5322 16805 13437 173025 ...
##  $ popdensity          : num  1271 759 681 1812 324 ...
##  $ popwhite            : int  63917 7054 14477 29344 5264 35157 5298 16519 13384 146506 ...
##  $ popblack            : int  1702 3496 429 127 547 50 1 111 16 16559 ...
##  $ popamerindian       : int  98 19 35 46 14 65 8 30 8 331 ...
##  $ popasian            : int  249 48 16 150 5 195 15 61 23 8033 ...
##  $ popother            : int  124 9 34 1139 6 221 0 84 6 1596 ...
##  $ percwhite           : num  96.7 66.4 96.6 95.3 90.2 ...
##  $ percblack           : num  2.575 32.9 2.862 0.412 9.373 ...
##  $ percamerindan       : num  0.148 0.179 0.233 0.149 0.24 ...
##  $ percasian           : num  0.3768 0.4517 0.1067 0.4869 0.0857 ...
##  $ percother           : num  0.1876 0.0847 0.2268 3.6973 0.1028 ...
##  $ popadults           : int  43298 6724 9669 19272 3979 23444 3583 11323 8825 95971 ...
##  $ perchsd             : num  75.1 59.7 69.3 75.5 68.9 ...
```

```
##  $ percollege          : num  19.6 11.2 17 17.3 14.5 ...
##  $ percprof            : num  4.36 2.87 4.49 4.2 3.37 ...
##  $ poppovertyknown     : int  63628 10529 14235 30337 4815 35107 5241 16455 13081 154934 ...
##  $ percpovertyknown    : num  96.3 99.1 95 98.5 82.5 ...
##  $ percbelowpoverty    : num  13.15 32.24 12.07 7.21 13.52 ...
##  $ percchildbelowpovert: num  18 45.8 14 11.2 13 ...
##  $ percadultpoverty    : num  11.01 27.39 10.85 5.54 11.14 ...
##  $ percelderlypoverty  : num  12.44 25.23 12.7 6.22 19.2 ...
##  $ inmetro             : int  0 0 0 1 0 0 0 0 0 1 ...
##  $ category            : chr  "AAR" "LHR" "AAR" "ALU" ...
```

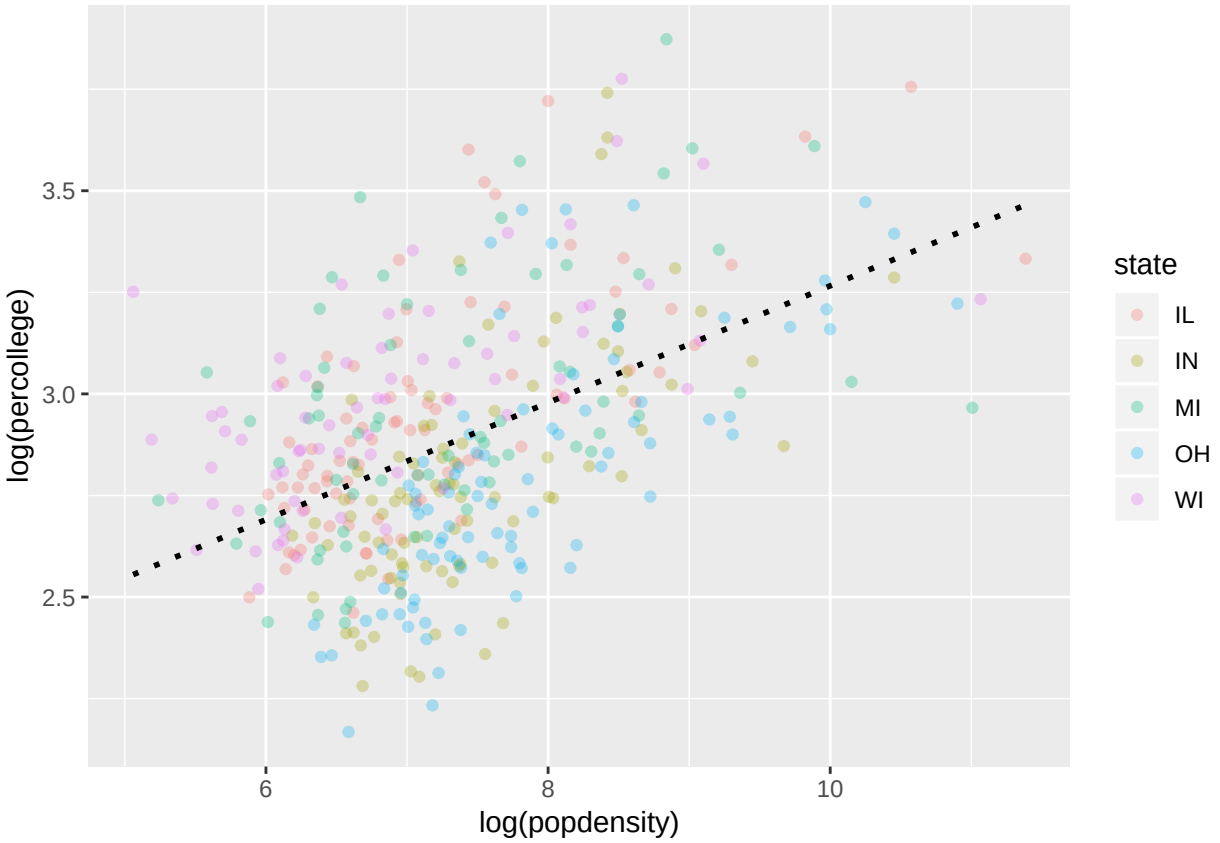
Q3: The plot of total population in each state.

```
midwest %>%
  group_by(state) %>%
  transmute(sum_state_pop = sum(poptotal, na.rm = T)) %>%
  ggplot(midwest, mapping = aes(x = sum_state_pop, fill = state)) +
  geom_histogram(bins = 10, alpha = 0.3, position = 'identity') +
  scale_x_continuous(labels = comma)
```



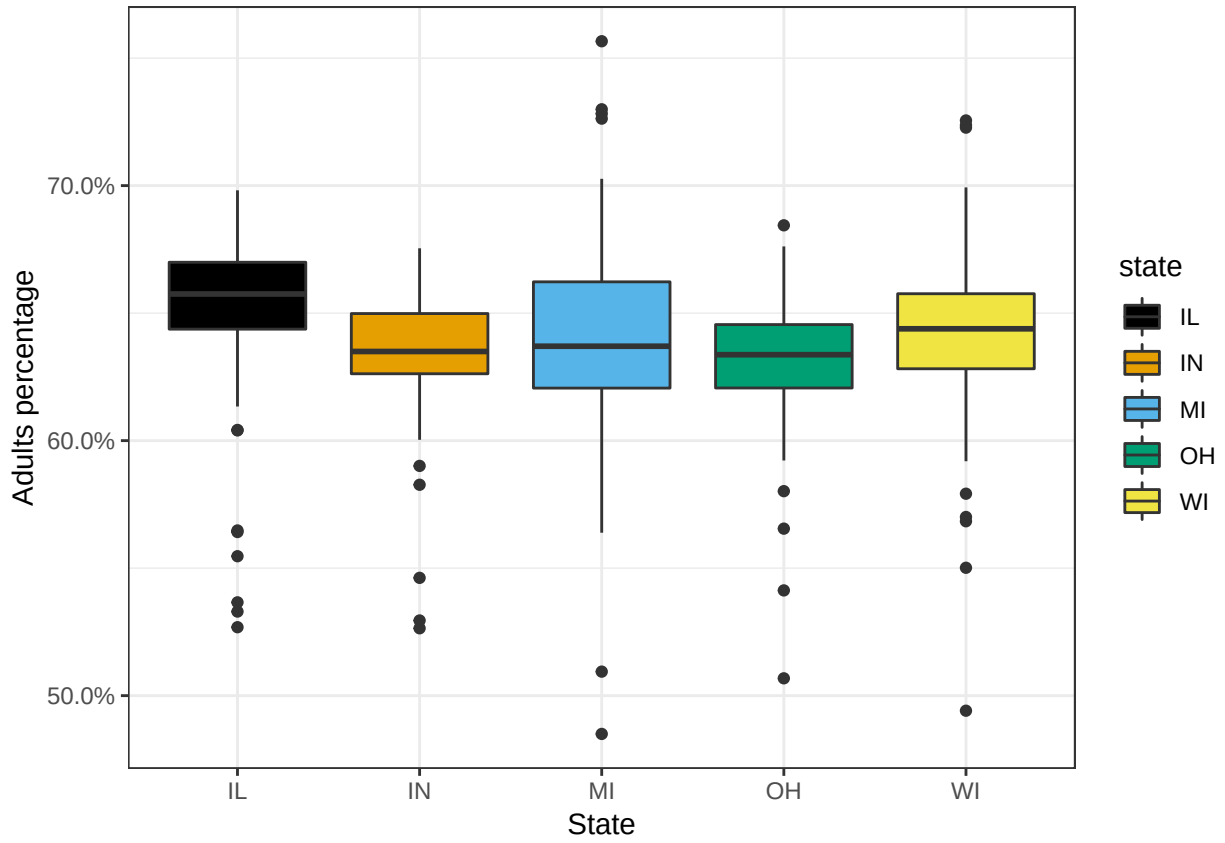
Q4: First find out the baseline value of the bottom tenth percentile of total population by using 'sort' function.The 44th 'poptotal' value is 12147.

```
new_midwest <- subset(midwest, poptotal > 12147)
ggplot(new_midwest, mapping = aes(x = log(popdensity), y = log(percollege), color = state)) +
  geom_point(alpha = 0.3) +
  geom_smooth(se = F, method = lm, linetype = 'dotted', color = 'black')
```

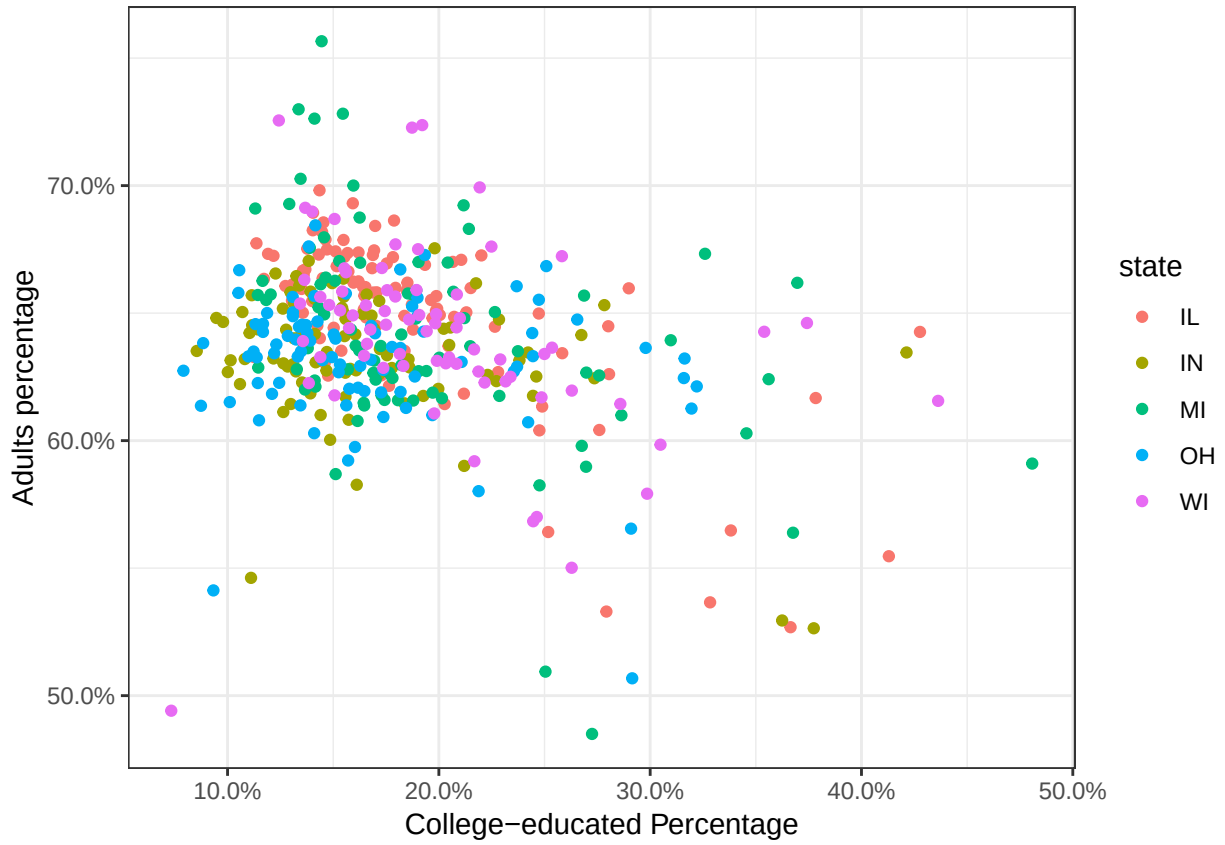


Q5: Overall, State Illinois and State Wisconsin have a higher adults percentage than other states. When taking account to 'percent of college-educated', it seems like there are no significant difference between each state.

```
midwest %>%
  group_by(state) %>%
  mutate(peradults = popadults / poptotal) %>%
  ggplot(midwest, mapping = aes(x = state, y = peradults, fill = state)) +
    geom_boxplot() +
    theme_bw() +
    scale_fill_colorblind() +
    scale_y_continuous('Adults percentage', labels = percent) +
    xlab('State')
```




```
midwest %>%
  group_by(state) %>%
  mutate(peradults = popadults / poptotal) %>%
  ggplot(midwest, mapping = aes(x = percollege/100, y = peradults, color= state)) +
    geom_point() +
    theme_bw() +
    scale_x_continuous('College-educated Percentage', label = percent) +
    scale_y_continuous('Adults percentage', labels = percent)
```



Q6: To determine the possible values of state.

```
unique(midwest$state)
```

```
## [1] "IL" "IN" "MI" "OH" "WI"
```

Q7: Replace state variable's value name.

```
midwest$state[midwest$state == "IL"] <- 'Illinois'
midwest$state[midwest$state == 'IN'] <- 'Indiana'
midwest$state[midwest$state == 'MI'] <- 'Michigan'
midwest$state[midwest$state == 'OH'] <- 'Ohio'
midwest$state[midwest$state == 'WI'] <- 'Wisconsin'
unique(midwest$state)
```

```
## [1] "Illinois"  "Indiana"   "Michigan"  "Ohio"      "Wisconsin"
```