DA5020.A2.Hsiao-Yu.Peng

Hsiao-Yu Peng

2023-09-13

R Markdown

Q1-1 See the attached certificate.

Q1-2

```
# load msleep dataset
data("msleep")

# look over msleep dataset
str(msleep)
```

```
## tibble [83 x 11] (S3: tbl_df/tbl/data.frame)
##  $ name        : chr [1:83] "Cheetah" "Owl monkey" "Mountain beaver" "Greater short-tailed shrew" ..
##  $ genus       : chr [1:83] "Acinonyx" "Aotus" "Aplodontia" "Blarina" ...
##  $ vore        : chr [1:83] "carni" "omni" "herbi" "omni" ...
##  $ order       : chr [1:83] "Carnivora" "Primates" "Rodentia" "Soricomorpha" ...
##  $ conservation: chr [1:83] "lc" NA "nt" "lc" ...
##  $ sleep_total : num [1:83] 12.1 17 14.4 14.9 4 14.4 8.7 7 10.1 3 ...
##  $ sleep_rem   : num [1:83] NA 1.8 2.4 2.3 0.7 2.2 1.4 NA 2.9 NA ...
##  $ sleep_cycle : num [1:83] NA NA NA 0.133 0.667 ...
##  $ awake       : num [1:83] 11.9 7 9.6 9.1 20 9.6 15.3 17 13.9 21 ...
##  $ brainwt     : num [1:83] NA 0.0155 NA 0.00029 0.423 NA NA NA 0.07 0.0982 ...
##  $ bodywt      : num [1:83] 50 0.48 1.35 0.019 600 ...
```

```
# if any missing value
any(is.na(msleep))
```

```
## [1] TRUE
```

The dataset "msleep" has dimension of 83 rows and 11 columns. It includes missing values within the dataset. The variables in the dataset encompass both character and numerics data types.

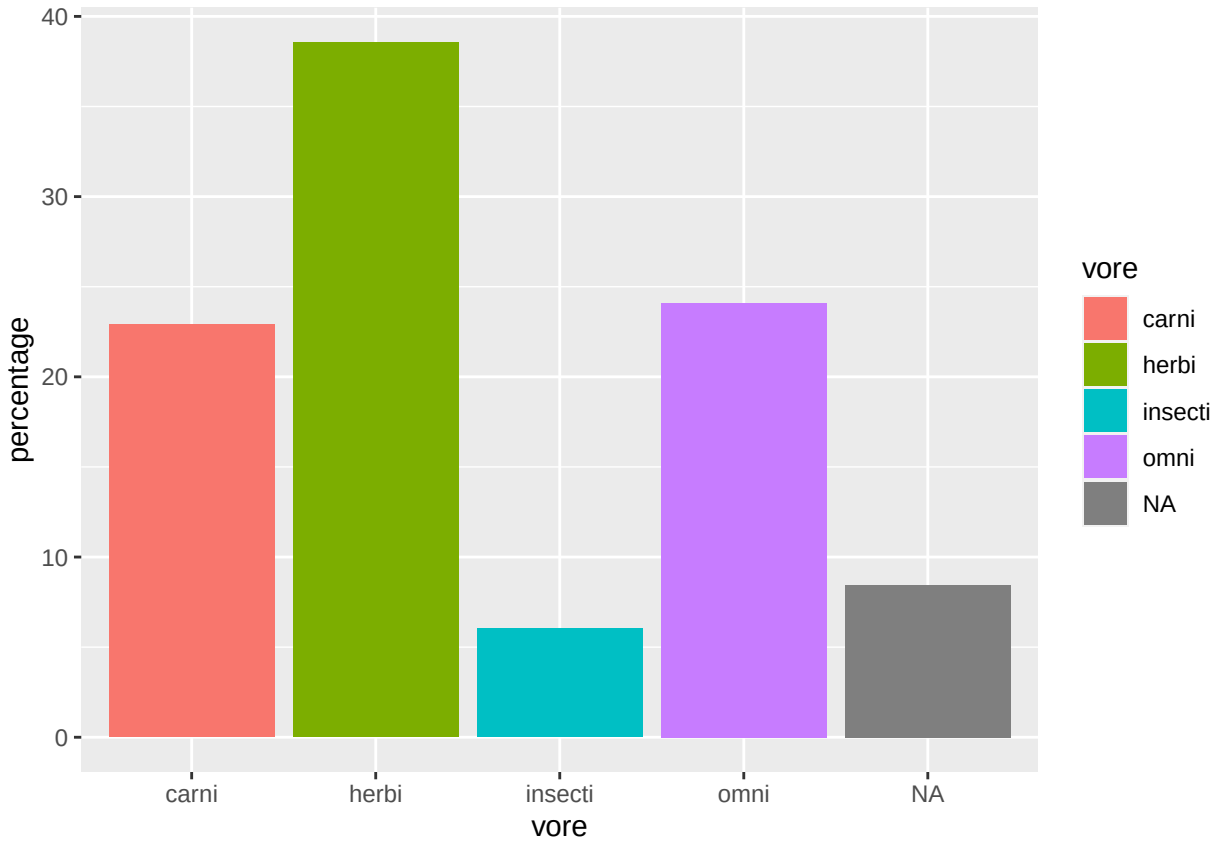
Q2.

```r
# find out the pecentage of each type of vore
vore_count <- msleep %>%
  group_by(vore) %>%
  summarize(n=n()) %>%
  mutate(percentage = n/sum(n) * 100)

print(vore_count)
```

```
## # A tibble: 5 x 3
##   vore        n percentage
##   <chr>   <int>      <dbl>
## 1 carni      19      22.9
## 2 herbi      32      38.6
## 3 insecti     5       6.02
## 4 omni       20      24.1
## 5 <NA>        7       8.43
```

```r
# Visualization for bar chart
ggplot(vore_count, aes(vore, percentage, fill = vore)) +
  geom_bar(stat = "identity")
```



Q3.

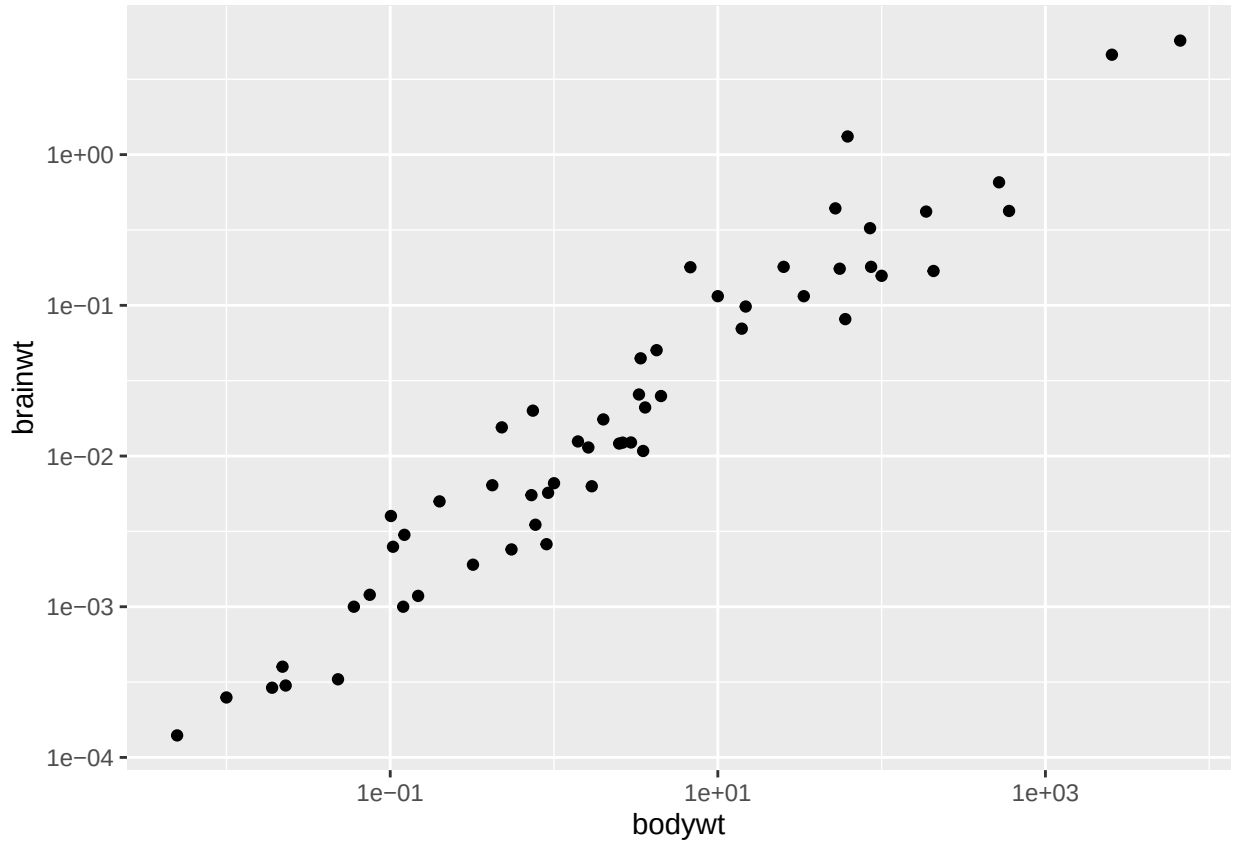
```
# Extract data for omnivores. Find its mean sleep_total
msleep %>%
  filter(vore =='omni') %>%
  summarise(mean(sleep_total))
```

```
## # A tibble: 1 x 1
##   `mean(sleep_total)`
##                 <dbl>
## 1                10.9
```

Q4

```
# Show a scatterplot between bodywt and brainwt
ggplot(msleep) + geom_point(aes(bodywt, brainwt)) +
  scale_x_log10() +
  scale_y_log10()
```

```
## Warning: Removed 27 rows containing missing values (`geom_point()`).
```



The scatterplot indicates that there may be a positive correlation between body with (bodywt) and brain weight (brainwt).

Q5.

```
# transform bodywt, brainwt
log_bw <- log10(msleep$bodywt)
log_brainwt <- log10(msleep$brainwt)

# Calculate Pearson's coefficient
# ignore any NA values: `use = "complete.obs"`
correlation <- cor(log_bw, log_brainwt, use = "complete.obs")

# result
correlation
```

```
## [1] 0.9653246
```

Yes, the Pearson coefficient of correlation is 0.965, which supports the assumption made in question 4.

Q6.

```
# Calculate mean and standard deviation of sleep_total
mean_sleep_total <- mean(msleep$sleep_total)
sd_sleep_total <- sd(msleep$sleep_total)

# Identify outliers
outliers <- msleep %>%
  filter(abs(sleep_total - mean_sleep_total) > 1.5 * sd_sleep_total)

# Display the name and sleep_total of outliers
select(outliers, name, sleep_total)
```

```
## # A tibble: 13 x 2
##    name                   sleep_total
##    <chr>                        <dbl>
##  1 Roe deer                       3
##  2 Long-nosed armadillo          17.4
##  3 North American Opossum        18
##  4 Big brown bat                 19.7
##  5 Horse                          2.9
##  6 Donkey                         3.1
##  7 Giraffe                        1.9
##  8 Pilot whale                    2.7
##  9 African elephant               3.3
## 10 Thick-tailed opposum          19.4
## 11 Little brown bat              19.9
## 12 Caspian seal                   3.5
## 13 Giant armadillo               18.1
```