msleep dataset visualization

Kanyaphat Lokhan

2025-06-05

Hello! This project is homework for the DataRockie Bootcamp #11, specifically the data visualization part. I've been practicing my visualization skills in R for a while, and this time I'm using the msleep dataset from ggplot2 to create some visualizations.

1. Data Overview

First of all, let's have a look at the data structure.

```
library(dplyr)
library(ggplot2)
df <- msleep
glimpse(df)
```

```
## Rows: 83
## Columns: 11
## $ name         <chr> "Cheetah", "Owl monkey", "Mountain beaver", "Greater shor~
## $ genus        <chr> "Acinonyx", "Aotus", "Aplodontia", "Blarina", "Bos", "Bra~
## $ vore         <chr> "carni", "omni", "herbi", "omni", "herbi", "herbi", "carn~
## $ order        <chr> "Carnivora", "Primates", "Rodentia", "Soricomorpha", "Art~
## $ conservation <chr> "lc", NA, "nt", "lc", "domesticated", NA, "vu", NA, "dome~
## $ sleep_total  <dbl> 12.1, 17.0, 14.4, 14.9, 4.0, 14.4, 8.7, 7.0, 10.1, 3.0, 5~
## $ sleep_rem    <dbl> NA, 1.8, 2.4, 2.3, 0.7, 2.2, 1.4, NA, 2.9, NA, 0.6, 0.8, ~
## $ sleep_cycle  <dbl> NA, NA, NA, 0.1333333, 0.6666667, 0.7666667, 0.3833333, N~
## $ awake        <dbl> 11.9, 7.0, 9.6, 9.1, 20.0, 9.6, 15.3, 17.0, 13.9, 21.0, 1~
## $ brainwt      <dbl> NA, 0.01550, NA, 0.00029, 0.42300, NA, NA, NA, 0.07000, 0~
## $ bodywt       <dbl> 50.000, 0.480, 1.350, 0.019, 600.000, 3.850, 20.490, 0.04~
```

The data has 83 rows and 11 columns. I'm going to use these columns building 5 charts. Let's get started!

2. Building Chart

(1) Number of animals by diet type

In this chart I'm gonna find the number of animals by diet type or the vore column, so I have to check first if there's any missing value.

```
df %>%
  filter(is.na(vore))
```

```
## # A tibble: 7 x 11
##   name    genus vore  order conservation sleep_total sleep_rem sleep_cycle awake
##   <chr>   <chr> <chr> <chr> <chr>              <dbl>     <dbl>       <dbl> <dbl>
## 1 Vesper~ Calo~ <NA>  Rode~ <NA>                 7        NA        NA      17
## 2 Desert~ Para~ <NA>  Erin~ lc                  10.3       2.7      NA      13.7
## 3 Deer m~ Pero~ <NA>  Rode~ <NA>                11.5      NA        NA      12.5
```

```
## 4 Phalan~ Phal~ <NA>  Dipr~ <NA>                13.7       1.8      NA      10.3
## 5 Rock h~ Proc~ <NA>  Hyra~ lc                   5.4       0.5      NA      18.6
## 6 Mole r~ Spal~ <NA>  Rode~ <NA>                10.6       2.4      NA      13.4
## 7 Musk s~ Sunc~ <NA>  Sori~ <NA>                12.8       2         0.183  11.2
## # i 2 more variables: brainwt <dbl>, bodywt <dbl>
```

As you can see, there are 7 missing values in vore column, I decided to search and add values in each row manually because there are less missing values.

```
clean_vore_data <- df %>%
  mutate(
    vore = case_when(
      name == 'Vesper mouse' ~ 'omni',
      name == 'Desert hedgehog' ~ 'omni',
      name == 'Deer mouse' ~ 'omni',
      name == 'Phalanger' ~ 'herbi',
      name == 'Rock hyrax' ~ 'herbi',
      name == 'Mole rat' ~ 'herbi',
      name == 'Musk shrew' ~ 'insecti',
      TRUE ~ vore
    )
  )
```

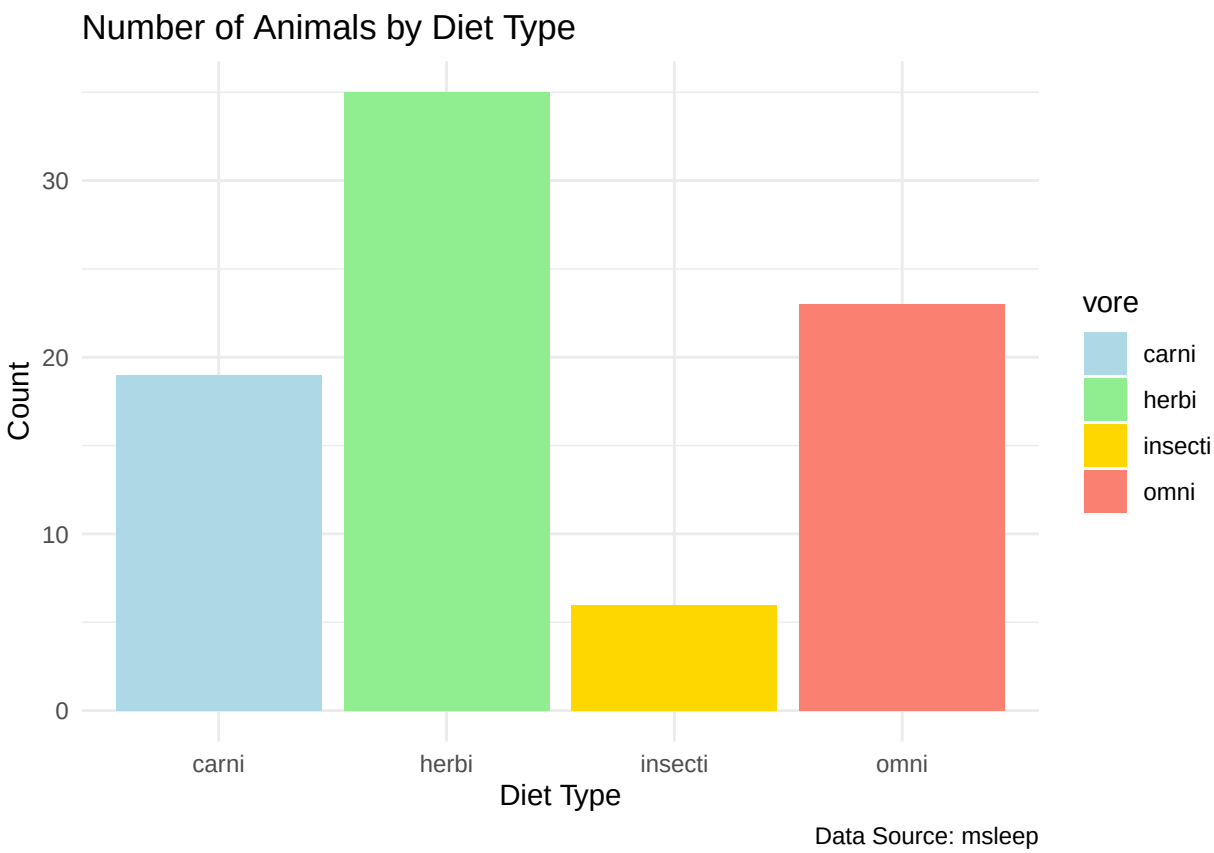
The last line (TRUE ~ vore) means to keep the values of the animal name I didn't mention above as the same. Then, I assigned this code to the new variable and checked again to make sure that all of the missing values gone.

```
clean_vore_data %>%
  filter(is.na(vore))
```

```
## # A tibble: 0 x 11
## # i 11 variables: name <chr>, genus <chr>, vore <chr>, order <chr>,
## #   conservation <chr>, sleep_total <dbl>, sleep_rem <dbl>, sleep_cycle <dbl>,
## #   awake <dbl>, brainwt <dbl>, bodywt <dbl>
```

After this process, the chart is ready to be built! In this case, I choose building bar chart because x (diet type) is a categorial data, and y normally is a stats, which is number of animal (n).

```
number_by_vore <- clean_vore_data %>%
  ggplot(aes(vore, fill = vore)) +
  geom_bar() +
  labs(title = 'Number of Animals by Diet Type',
       x = 'Diet Type',
       y = 'Count',
       caption = 'Data Source: msleep') +
  theme_minimal() +
  scale_fill_manual(values = c('lightblue',
                               'lightgreen',
                               'gold',
                               'salmon'))

print(number_by_vore)
```

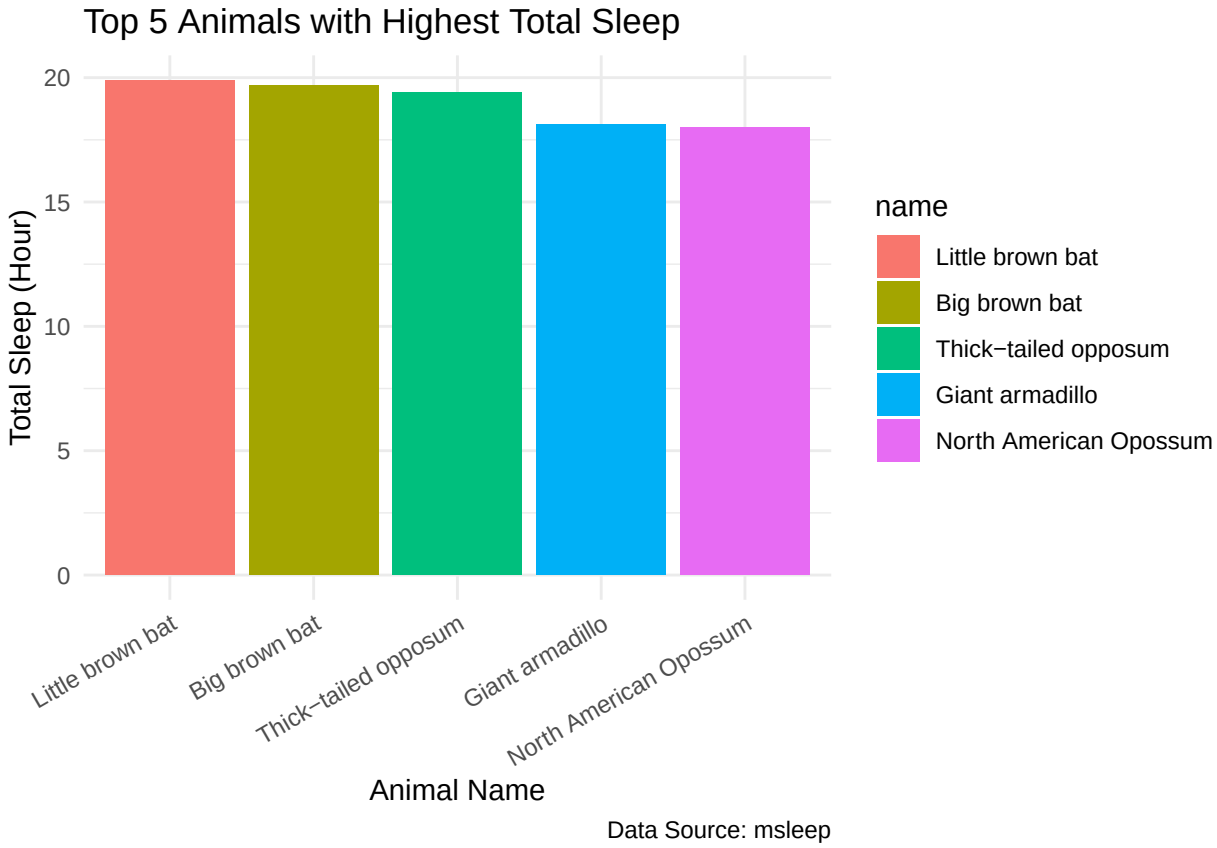


Herbivores have the highest number of animals with 35, followed by omnivores with 23. Carnivores and insectivores have fewer, with 19 and 6 respectively.

(2) Top 5 Animals with Highest Total Sleep

In this chart, I use dplyr preparing data first, by finding which first five animals has the highest total time of sleep, I use arrange to sort data by descending and limit only top five. I apply library(forcats) to use fct_reorder() function, which make the bars in chart order by their height.Then I build chart using ggplot.

```
library(forcats)

top_5_most_sleep <- clean_vore_data %>%
  arrange(desc(sleep_total)) %>%
  head(5) %>%
  mutate(name = fct_reorder(name, sleep_total, .desc = TRUE)) %>%
  ggplot(aes(x = name, y = sleep_total, fill = name)) +
  geom_col() +
  theme_minimal() +
  labs(title = 'Top 5 Animals with Highest Total Sleep',
       x = 'Animal Name',
       y = 'Total Sleep (Hour)',
       caption = 'Data Source: msleep') +
  theme(axis.text.x = element_text(hjust = 1, angle = 30))

print(top_5_most_sleep)
```

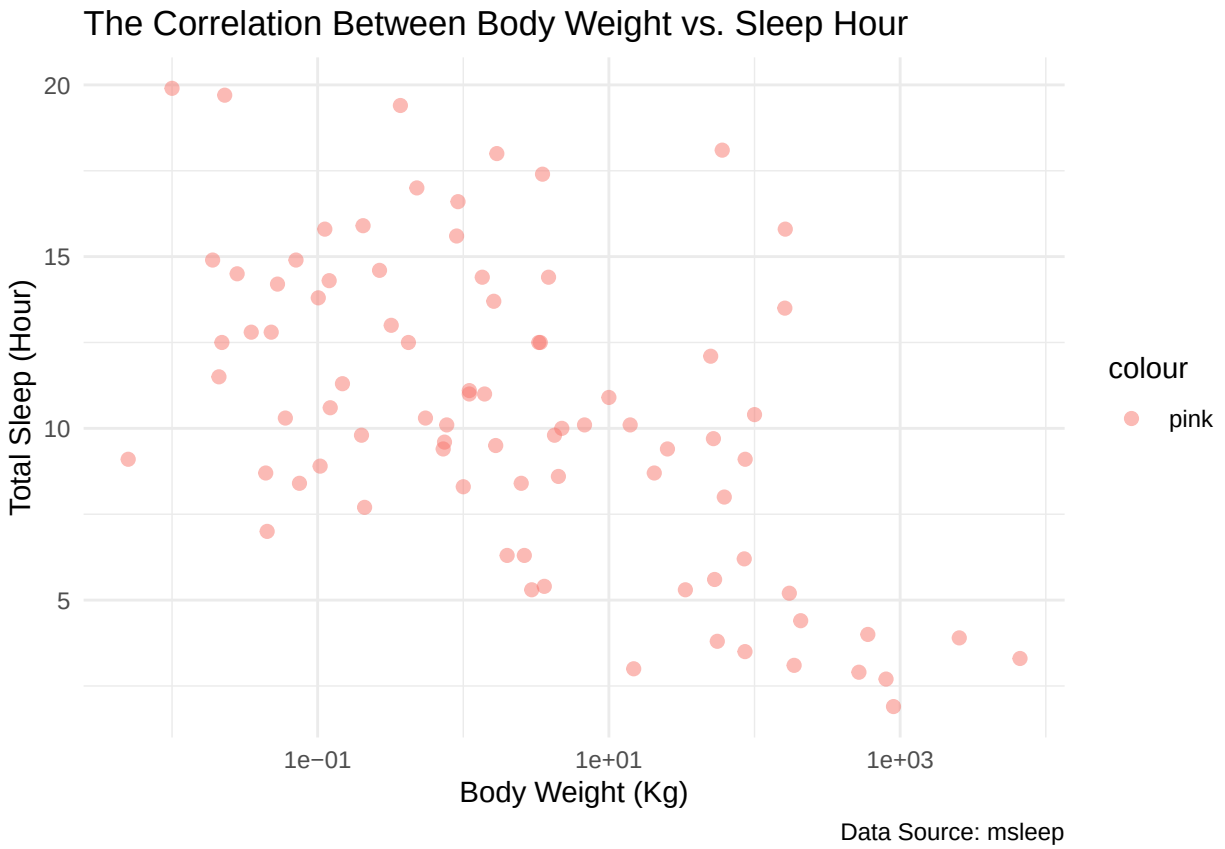


From the chart, we can see that the little brown bat has the most sleep time per day at 19.9 hours. It's followed by the big brown bat (19.7 hr/day), the thick-tailed opossum (19.4 hr/day), and the giant armadillo and North American opossum (18.1 and 18 hr/day, respectively). Since the first two animals are bats from the order Chiroptera, we can assume that bats are nocturnal animals and conserve energy during the day.

(3) The Correlation Between Body Weight vs. Sleep Hour

I plotted a scatter chart to explore the relationship between body weight and sleep hours. Due to the wide range of body weights, I transformed the x-axis to a log base 10 scale using the scale_x_log10() function. This method converts the numbers into scientific notation, making the chart more readable.

```
bodywt_totl_sleep <- ggplot(clean_vore_data,
       aes(x = bodywt, y = sleep_total)) +
  geom_point(aes(color = 'pink'), alpha = 0.5, size = 2) +
  scale_x_log10() +
  theme_minimal() +
  labs(title = 'The Correlation Between Body Weight vs. Sleep Hour',
       x = 'Body Weight (Kg)',
       y = 'Total Sleep (Hour)',
       caption = 'Data Source: msleep')

print(bodywt_totl_sleep)
```

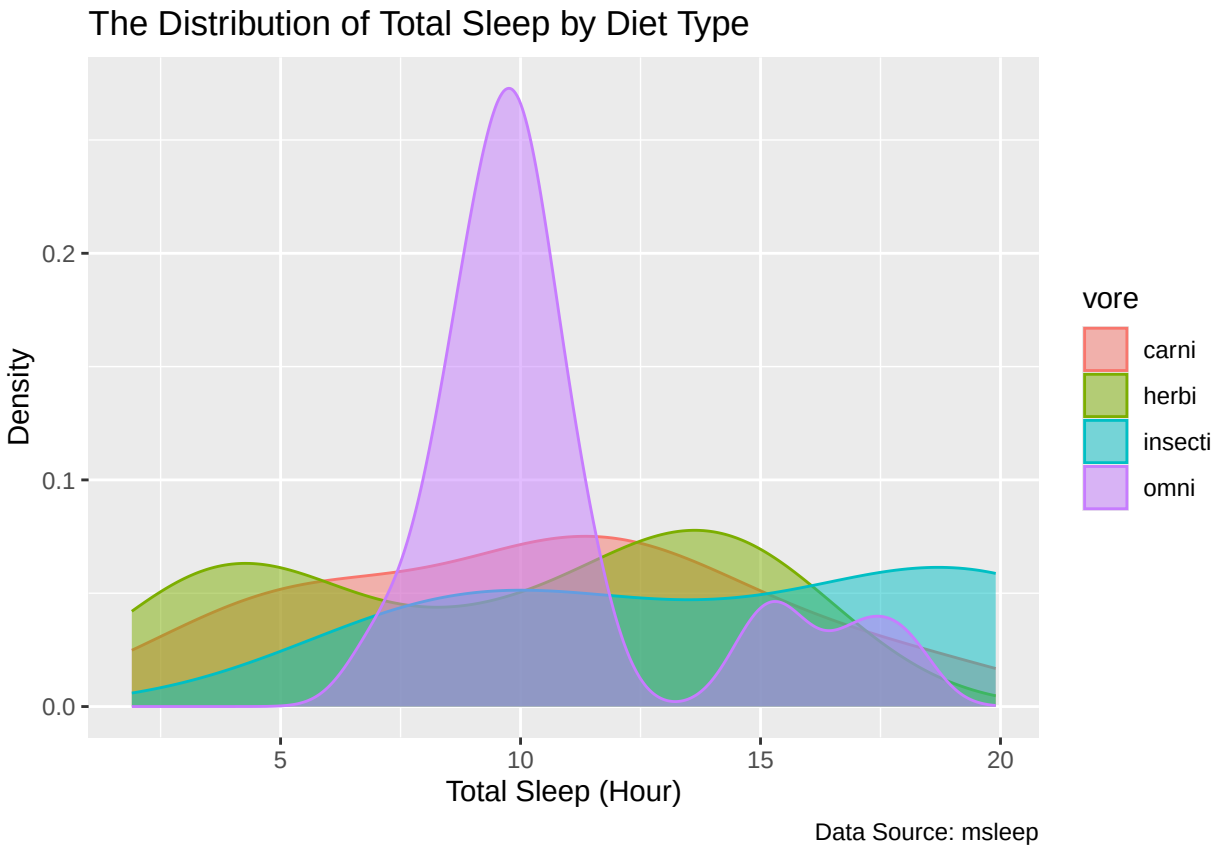


The chart shows that body weight tends to affect sleep time: heavier animals generally require less sleep.

(4) Distribution of Total Sleep by Vore Type

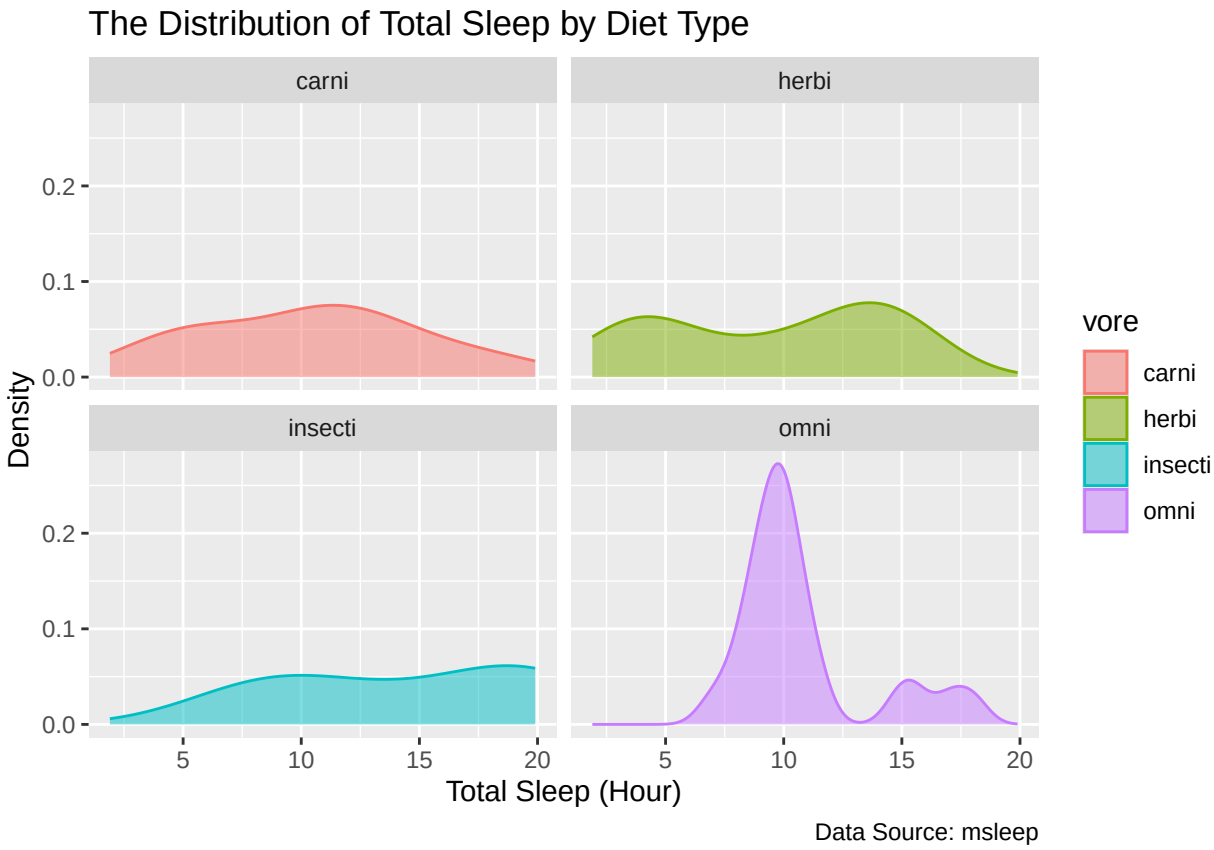
To investigate if diet type relates to sleep time, I plotted density chart to look at the distribution. This chart requires x in continuous type as an input so I use sleep_total here. I filled each density line in the chart with the diet type for the good-looking and readability.

```
totl_sleep_by_vore <- ggplot(clean_vore_data,
       aes(sleep_total, fill = vore, color = vore)) +
  geom_density(alpha = 0.5) +
  labs(title = 'The Distribution of Total Sleep by Diet Type',
       x = 'Total Sleep (Hour)',
       y = 'Density',
       caption = 'Data Source: msleep')

print(totl_sleep_by_vore)
```



From the plot, we can see a significant variation in total sleep duration among different diet types. While all diet types share some overlap in their sleep patterns, omnivores stand out with the highest distribution, showing a strong concentration of sleep around 10 hours. The other diet categories, however, display a flatter distribution. This overlap suggests that a mammal's diet alone isn't the sole determinant of its sleep duration. In this case, the density lines can be separated into small parts by using facet technique which makes the chart easily compare.

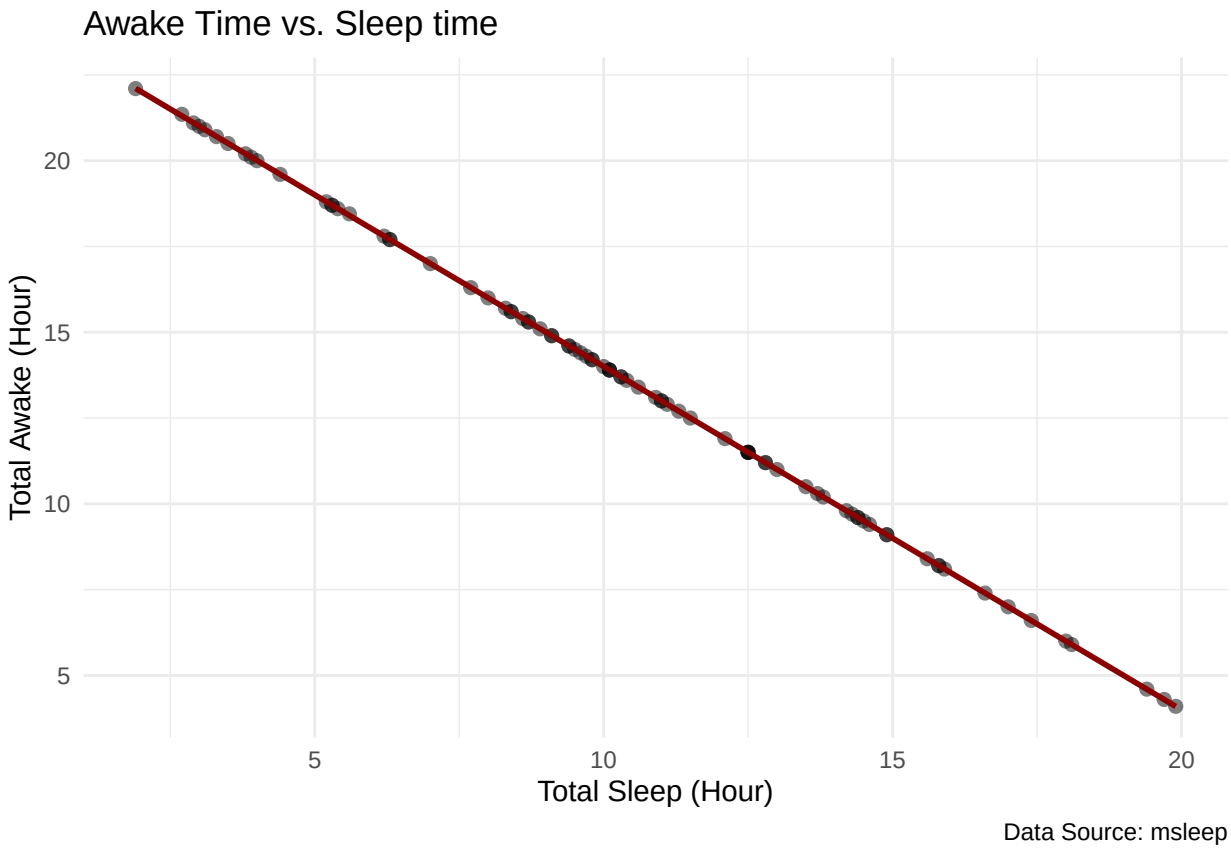
```
totl_sleep_by_vore_2 <- ggplot(clean_vore_data,
       aes(sleep_total, fill = vore, color = vore)) +
  geom_density(alpha = 0.5) +
  facet_wrap(~vore, ncol = 2) +
  labs(title = 'The Distribution of Total Sleep by Diet Type',
       x = 'Total Sleep (Hour)',
       y = 'Density',
       caption = 'Data Source: msleep')

print(totl_sleep_by_vore_2)
```



(5) The correlation between sleep time vs. awake time

We all know that these two columns have a negative correlation: when sleep time is high, awake time will be low, and vice-versa. I plotted a scatter chart with a trendline in linear regression model to clearly visualize this relationship.

```
awake_sleep <- ggplot(clean_vore_data,
       aes(sleep_total, awake)) +
  geom_point(alpha = 0.5, size = 2) +
  geom_smooth(method = 'lm',
              se = TRUE,
              color = 'darkred') +
  theme_minimal() +
  labs(title = 'Awake Time vs. Sleep time',
       x = 'Total Sleep (Hour)',
       y = 'Total Awake (Hour)',
       caption = 'Data Source: msleep')

print(awake_sleep)
```



Project Ends