Lab 2 – Mammal Sleep Behavior

NAME 1 – varenya3

NAME 2 – NETID [if applicable]

NAME 3 – NETID [if applicable]

Formatting Instructions

- Please include all requested responses in a document, then save it as a **pdf** when done.
  - You may use this instructions document, or you may create a new document.
  - All responses should be numbered (leaving the original question text is optional!)
- Upload your pdf to **Gradescope** and please **match pages** with the **question number** when prompted to.
- If working with one or two **partners**, be sure to do **both** of these things:
  - Please put all **names and netIDs** at the top of your document (like shown above).
  - Have one person upload the pdf and then ensure **group members** are **added** in your submission to Gradescope (click view/edit group on the top right of the page once shown your final submission after matching pages).

Assignment Overview

- What is the sleep behavior of different mammals? How does it vary across species?
- You will have a chance to explore some data about this using R, and then briefly commenting on some of your findings.

Step 0

- **Pre-lab work**
  - Complete the pre-lab tutorials for Lab 2 first: https://stat212-learnr.stat.illinois.edu/
  - Watch videos 1 (or 2), 3, 4, and 5 in this playlist:
    https://www.youtube.com/playlist?list=PLTE0IJCCTM9ILfW8OaLqZd37G7X4WDtl-
- **Open RStudio** (or RStudio Cloud) to get started
  - Be careful **not** open up **R** (this icon with just R and a swirly thing on the left).
  - Open up **RStudio** (this icon with the blue circle on the right!).
- **Open the starter script** linked in the assignment description.
  - Be sure you are writing your code in the script. I do **NOT** recommend coding directly into the console (command line). The console doesn't allow you to edit your code across multiple lines and it will be way more frustrating!
- **Install and library tidyverse**
  - Write and run the following code: `install.packages("tidyverse")`
  - This will take a minute or two! Wait until the little stop sign disappears to proceed.
    - *Remember that this only needs to be installed once—**no need to reinstall it for each new lab assignment**.*
  - Next, you will want to run the following code: `library(tidyverse)`
    - Note that you will need to run this code every time you open RStudio again and use any tidyverse functions, like ggplot. *If you **close** RStudio and **open it again**, you'll need to library tidyverse again.*
- **Open the `msleep`** data by running the code: `View(msleep)`
- **Open the documentation** for `msleep` by running the code: `?msleep`

Question 1 (5pts): Create a histogram of the `sleep_total` variable (using `ggplot2`). Your histogram should:

- Choose a black border color
- Change the fill color to the color of your choice (*something besides black*)
- Have an appropriate title
- Have a set number of bins (don't use the default of 30—try doing a lower number).

Include the image of your histogram in your report *(not sure how to do that? Watch the*

Include your R code for this question *(copy the image, import the image, or selective screenshot)*
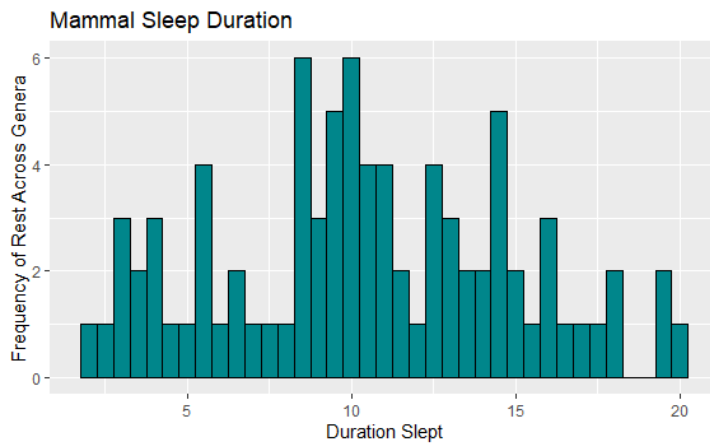
Mammal Sleep Duration

```
ggplot(data=msleep, aes(x=sleep_total))

+ geom_histogram(color='black', fill='turquoise4', bins=15, binwidth = 0.5)

+ labs(title = "Mammal Sleep Duration", x = "Duration Slept (Hours per Day)", y = "Frequency of Rest
Across Genera")
```

```
> ggplot(data=msleep, aes(x=sleep_total)) + geom_histogram(color='black', fill='turquoise
4', bins=15, binwidth = 0.5) + labs(title = "Mammal Sleep Duration", x = "Duration Slept
(Hours per Day)", y = "Frequency of Rest Across Genera")
```

Question 2 (5pts) Use the `summary` function to numerically summarize the `sleep_total` variable (and only this variable). *This function should output the min, Q1, Q2, mean, Q3, and max.*

> **Report the numeric summary**

> **Also report the standard deviation** of this same variable with a separate function.

> **Include your R code for this question**

```
> summary(msleep$sleep_total)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.90    7.85   10.10   10.43   13.75   19.90
> sd(msleep$sleep_total)
[1] 4.450357
```

Question 3 (5pts): Now **briefly describe the distribution of this variable**.

- What is a typical (median) amount of sleep for a mammal species in this dataset?
- What is lowest and highest amount of sleep recorded in this dataset?
- What is the average deviation from the mean (standard deviation) for total sleep in this dataset?
- In what sleep range are the middle 50% of mammal species in this dataset?

```
-   $ what is a typical (median) amount of sleep for a mammal species in this dataset?
-       > 10.10 Hours is the typical (median) amount of sleep for a mammal species.
-   $ what is lowest and highest amount of sleep recorded in this dataset?
-       > 1.90 and 19.90 hours are lowest and highest amount of sleep recorded.
-   $ what is the average deviation from the mean (standard deviation) for total sleep
    in this dataset?
-       > 4.450357 hours is the average deviation from the mean (standard deviation).
-   $ In what sleep range are the middle 50% of mammal species in this dataset?
-       > 5.90 is the IQR, or the sleep range of the middle 50% of mammal species.
```

Question 4 (5pts): Create a histogram of the body weight variable *(look at the data viewer to check exactly how this variable is named in the data frame!).* Your histogram should:

- Have a black border color
- Have the fill color of your choice (use a *different* color from your first one)
- Have an appropriate title
- Any number of bins is fine (at least 20, no more than 100)

Include the image of your histogram in your report

Include your R code for this question

Briefly describe how this distribution is different in comparison to the sleep total distribution.
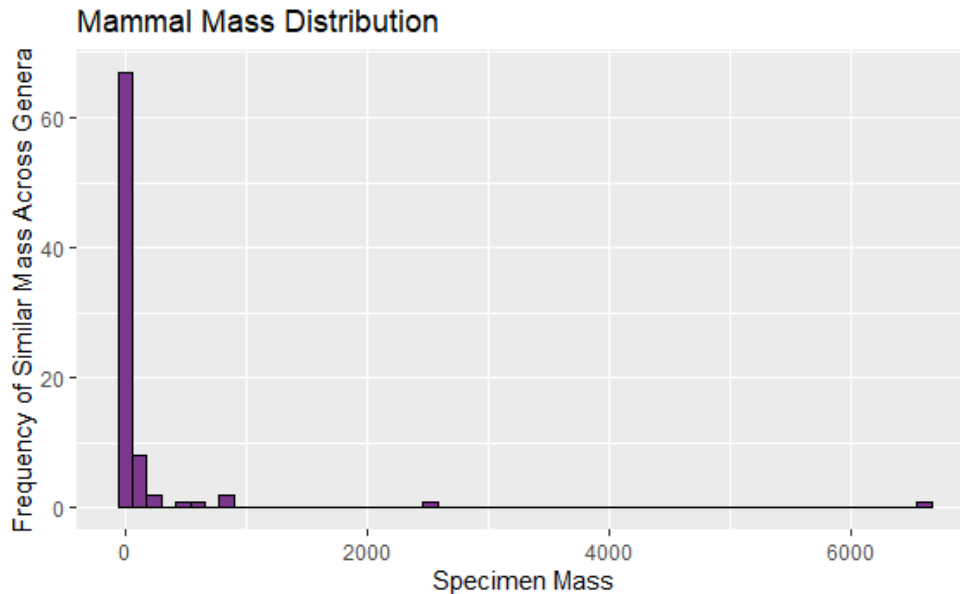
Mammal Mass Distribution

ggplot(data=msleep, aes(x=bodywt)) + geom_histogram(color='black', fill='mediumorchid4', bins=80, binwidth = 120) + labs(title = "Mammal Mass Distribution", x = "Specimen Mass (kg)", y = "Frequency of Similar Mass Across Genera")

```
-   > ggplot(data=msleep, aes(x=bodywt)) + geom_histogram(color='black', fill='mediumor
    chid4', bins=80, binwidth = 120) + labs(title = "Mammal Mass Distribution", x = "Sp
    ecimen Mass (kg)", y = "Frequency of Similar Mass Across Genera")
-
-   >This  distribution is different in comparison to the sleep total distribution due
    to the greater range and variance. as well as the more pronounced frequency of the
    measure variable in this case over the former.
```

Question 5 (5pts): What are the eating classifications for the mammals in this dataset? Let's answer this question by **creating a barplot** to compare the frequency of each `vore` type. Your barplot should:

- Have a black border color
- Allow each bar a different color (fill by the variable)
- Have an appropriate title

Include the image of your barplot in your report

Include your R code for this question

Which "vore" classification appears the most in this dataset?
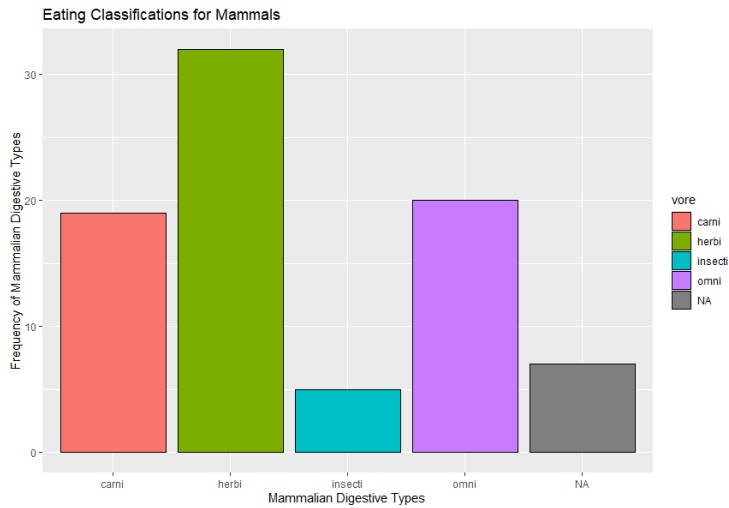
Eating Classifications for Mammals

ggplot(data=msleep, aes(x=vore), fill=vore) + geom_bar(colour='black', aes(fill=vore)) + labs(stat="identity", title = "Eating Classifications for Mammals", x = "Mammalian Digestive Types", y = "Frequency of Mammalian Digestive Types")
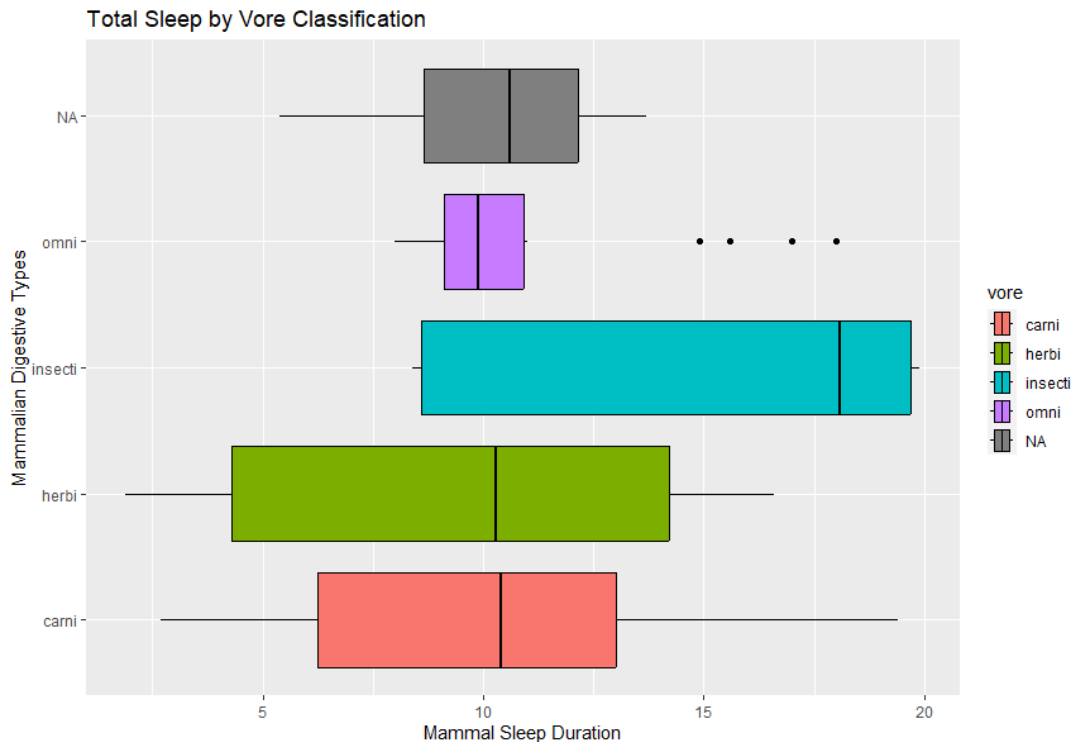
```
-   > ggplot(data=msleep, aes(x=vore), fill=vore) + geom_bar(colour='black', aes(fill=v
    ore)) + labs(stat="identity", title = "Eating Classifications for Mammals", x = "Ma
    mmalian Digestive Types", y = "Frequency of Mammalian Digestive Types")
-
-   > The Herbivore classification appears most in this dataset.
```

Question 6 (5pts): Create one graph containing side by side boxplots to compare total sleep by vore classification. You should have a separate boxplot for each of the 5 categories in vore (this will include an "NA" category). Your boxplots should:

-   Allow each box a different color (fill by the grouping variable)
-   Have an appropriate title
-   You can arrange them vertically or horizontally. Up to you!

Include the image of your side-by-side boxplots in your report

Include your R code for this question

Total Sleep by Vore Classification

```
> ggplot(data=msleep, aes(x=vore), fill=vore) + geom_bar(colour='black', aes(fill=v
ore)) + labs(stat="identity", title = "Eating Classifications for Mammals", x = "Ma
mmalian Digestive Types", y = "Frequency of Mammalian Digestive Types")
```

Question 7 (5pts): Answer these questions about the previous graph and by looking at the `msleep` data viewer more carefully. *Note, you can sort by a column by clicking on the column header!*

Which "vore" classification appears to get the most sleep in this dataset?

➢ The Insectivore classification of Mammals appears to get the most sleep in this dataset.

Which specific animals in that classification are getting the most? (check the data viewer and sort!)

➢ The Little brown bat of Genus Myotis under Order Chiroptera.

Is high total sleep a *consistent* feature of all animals in that classification, or would you say that these animals vary a lot?

➢ This animal classification has the highest variance out of all the others in this dataset.

In general, do you find "vore" classification to be a helpful way to explain variability in total sleep across mammals (in other words, does knowing a mammal's "vore" classification give us much indication of their total sleep)? **Why or why not?**

➢ Though the average sleep duration is higher than all the other vore types, the variance is over 50% higher than the next greatest "vore" type, and it has the least number of specimens in the dataset. Therefore we cannot certainly concluded that "vore" classification is a helpful way to explain variability in total sleep across mammals.

- $ Which "vore" classification appears to get the most sleep in this dataset?
- > The Insectivore classification of Mammals appears to get the most sleep in this dataset.
- > msleep %>% arrange(-sleep_total) %>% head()
-
- $ Which specific animals in that classification are getting the most? (check the data viewer and sort!)
- > The Little brown bat of Genus Myotis under Order Chiroptera.
- > msleep %>%
-     select(name, vore, sleep_total) %>%
-     arrange(-sleep_total, vore) %>%
-     head()
-
- $ Is high total sleep a consistent feature of all animals in that classification, or would you say that these animals vary a lot?
- > This animal classification has the highest variance out of all the others in this dataset.
- > Referring to Question 5, this are the least frequent classification of mammals.
- > aggregate(SleepingData$sleep_total, by=list(SleepingData$vore), FUN=var)
- > ggplot(data=msleep, aes(x=vore), fill=vore) + geom_bar(colour='black', aes(fill=vore)) + labs(stat="identity", title = "Eating Classifications for Mammals", x = "Mammalian Digestive Types", y = "Frequency of Mammalian Digestive Types")
-
- $ In general, do you find "vore" classification to be a helpful way to explain variability in total sleep across mammals (in other words, does knowing a mammal's "vore" classification give us much indication of their total sleep)? Why or why not?
- > Though the average sleep duration is higher than all the other vore types, the variance is over 50% higher than the next greatest "vore" type, and it has the least number of specimens in the dataset. Therefore, we cannot certainly concluded that "vore" classification is a helpful way to explain variability in total sleep across mammals.