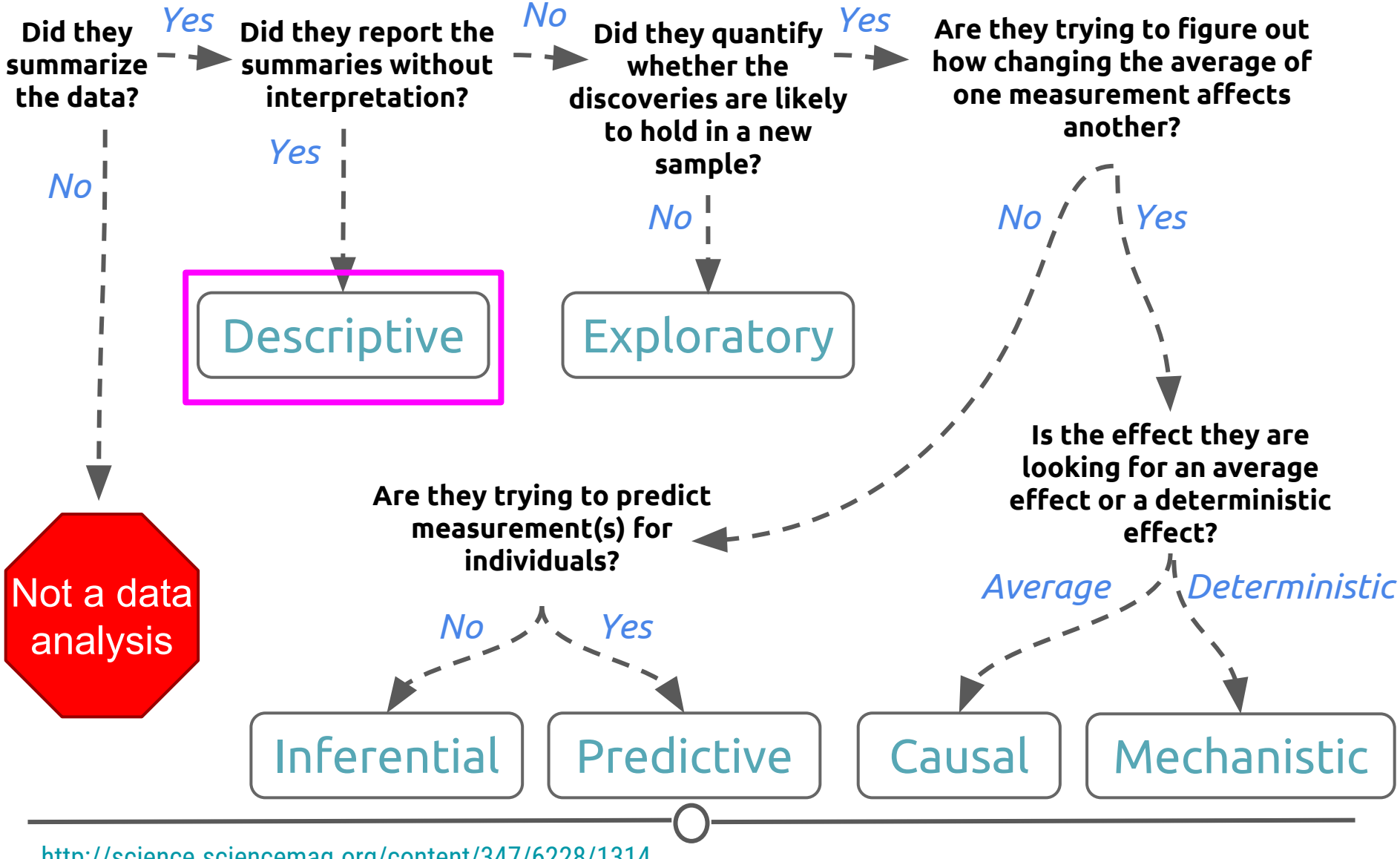
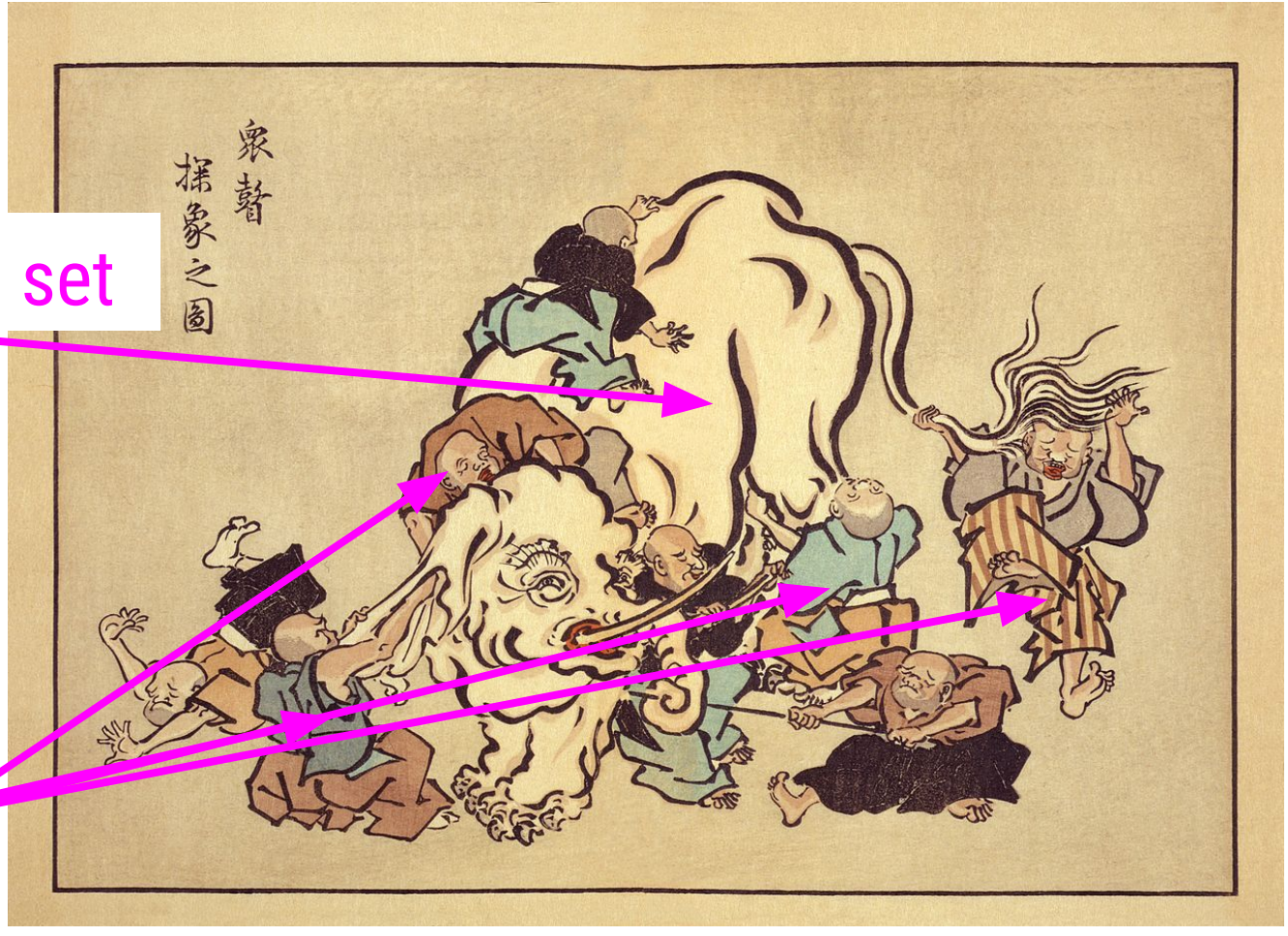
Descriptive Analysis

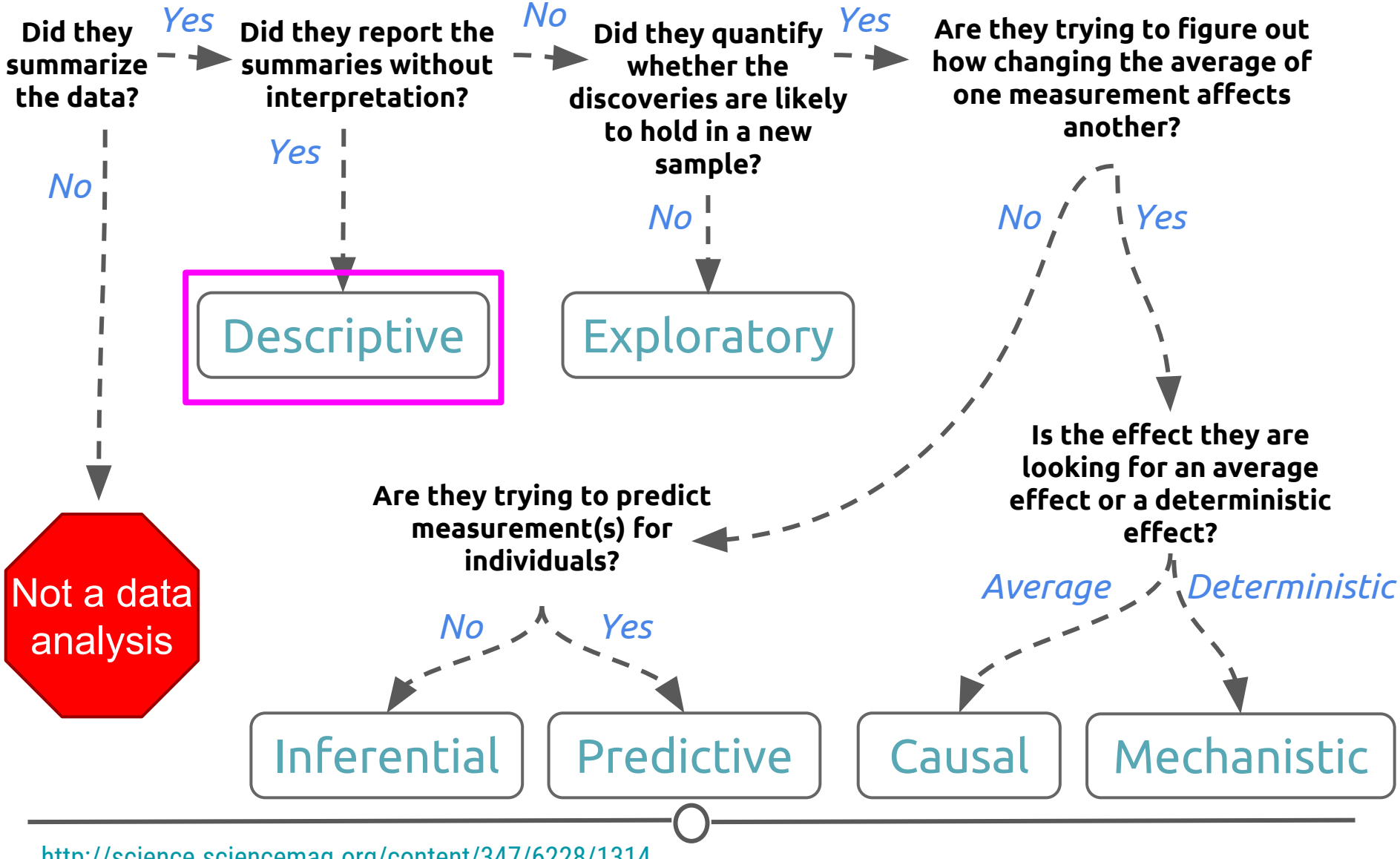
Data Analysis



http://science.sciencemag.org/content/347/6228/1314

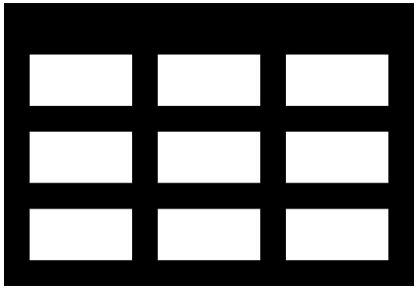
https://en.wikipedia.org/wiki/Blind_men_and_an_elephant#/media/File:Blind_monks_examining_an_elephant.jpg


衆瞽
摸象之圖
The data set
You



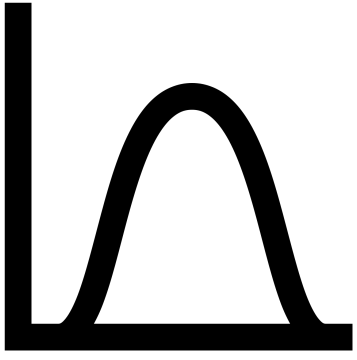
http://science.sciencemag.org/content/347/6228/1314

Descriptive Analysis

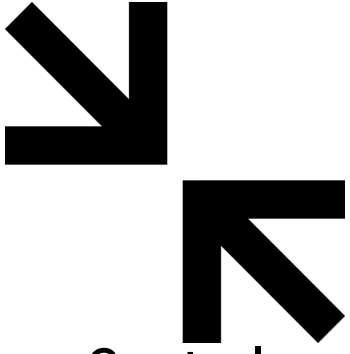


Size

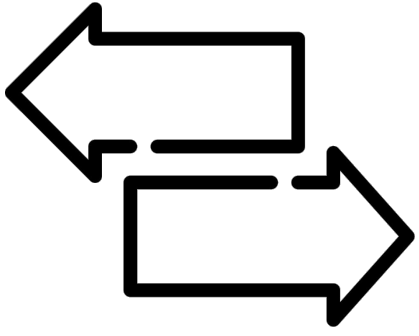
Missingness



Shape



Central Tendency



Variability

Subject	Total Estimate
Total population	309,349,689
AGE	
Under 5 years	6.5%
5 to 9 years	6.6%
10 to 14 years	6.7%
15 to 19 years	7.1%
20 to 24 years	7.0%
25 to 29 years	6.8%
30 to 34 years	6.5%
35 to 39 years	6.5%
40 to 44 years	6.8%
45 to 49 years	7.3%
50 to 54 years	7.2%
55 to 59 years	6.4%
60 to 64 years	5.5%
65 to 69 years	4.0%
70 to 74 years	3.0%
75 to 79 years	2.3%
80 to 84 years	1.9%
85 years and over	1.8%

2010 US Census Data Summary Table (broken down by age)

https://factfinder.census.gov/

Subject	United States		
	Total	Male	Female
	Estimate	Estimate	Estimate
Total population	309,349,689	152,089,450	157,260,239
AGE			
Under 5 years	6.5%	6.8%	6.3%
5 to 9 years	6.6%	6.8%	6.4%
10 to 14 years	6.7%	7.0%	6.4%
15 to 19 years	7.1%	7.5%	6.8%
20 to 24 years	7.0%	7.3%	6.7%
25 to 29 years	6.8%	6.9%	6.6%
30 to 34 years	6.5%	6.6%	6.4%
35 to 39 years	6.5%	6.6%	6.5%
40 to 44 years	6.8%	6.9%	6.7%
45 to 49 years	7.3%	7.3%	7.3%
50 to 54 years	7.2%	7.2%	7.2%
55 to 59 years	6.4%	6.3%	6.5%
60 to 64 years	5.5%	5.4%	5.6%
65 to 69 years	4.0%	3.9%	4.2%
70 to 74 years	3.0%	2.8%	3.2%
75 to 79 years	2.3%	2.1%	2.6%
80 to 84 years	1.9%	1.5%	2.2%
85 years and over	1.8%	1.2%	2.4%

... and stratified by sex

https://factfinder.census.gov/

An updated and expanded version of the mammals sleep dataset

Description

This is an updated and expanded version of the mammals sleep dataset. Updated sleep times and weights were taken from V. M. Savage and G. B. West. A quantitative, theoretical framework for understanding mammalian sleep. Proceedings of the National Academy of Sciences, 104 (3):1051-1056, 2007.

Usage

```
msleep
```

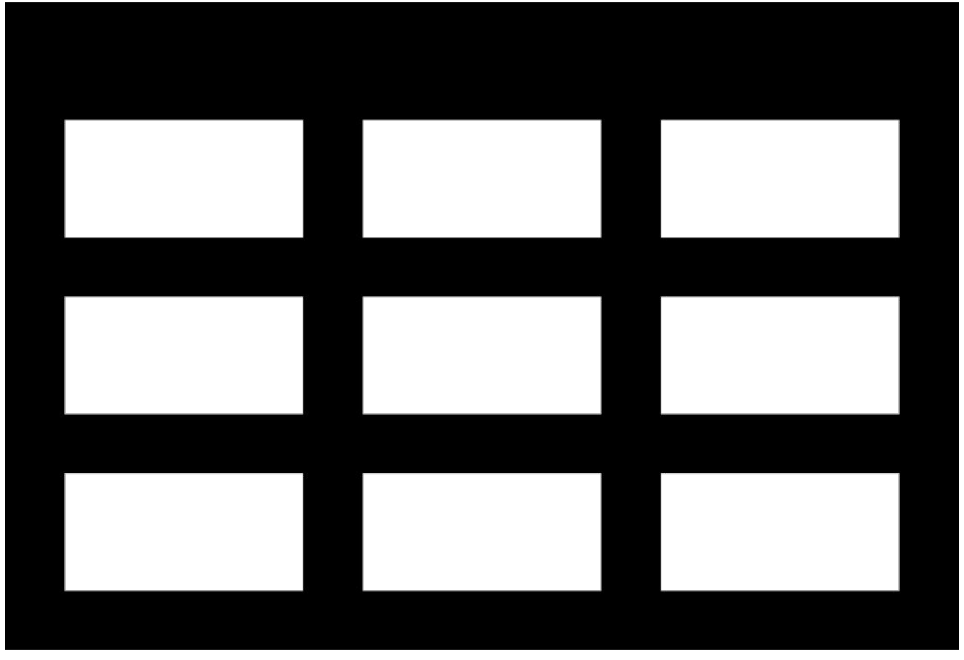
Format

A data frame with 83 rows and 11 variables

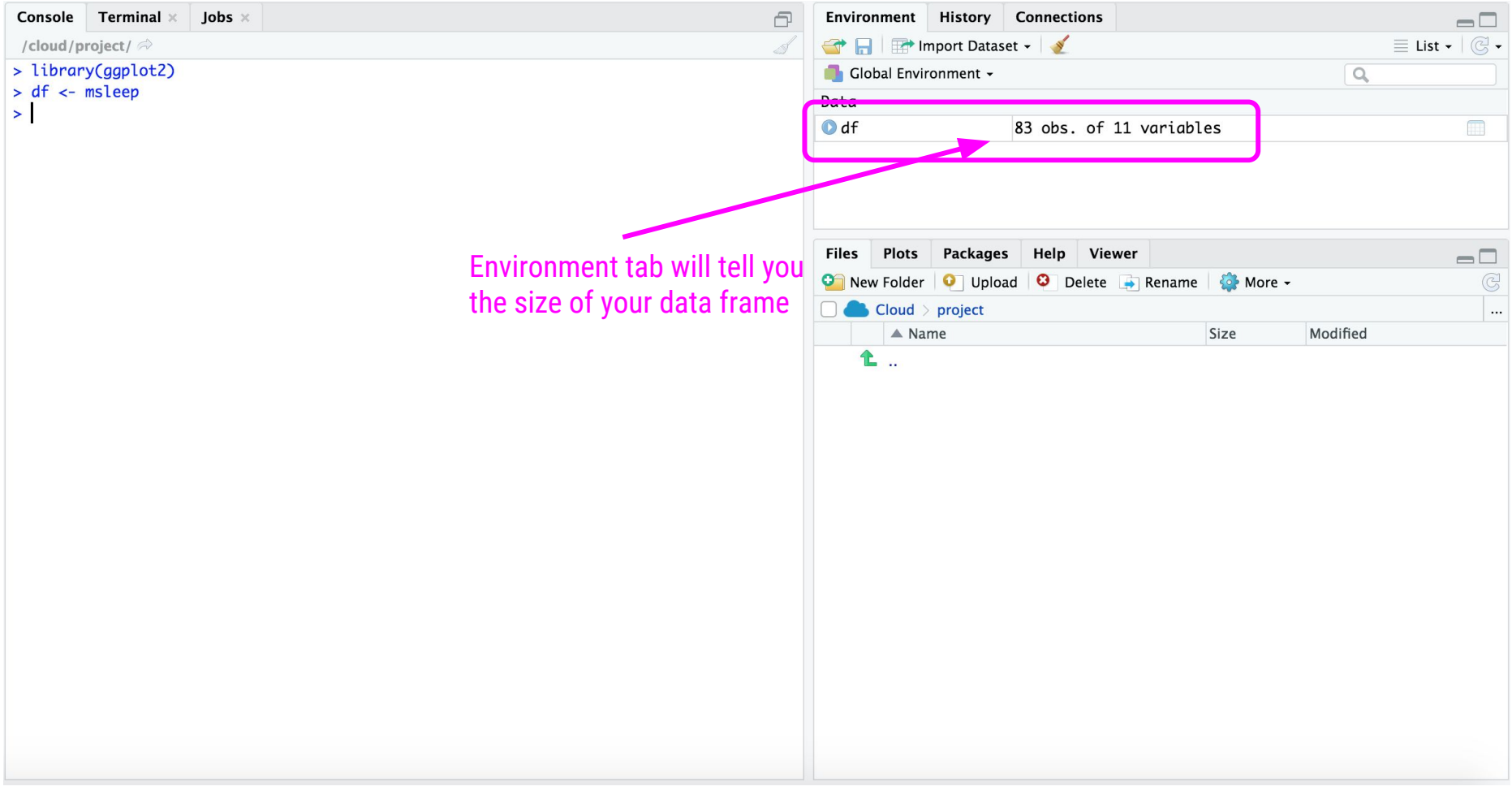
name

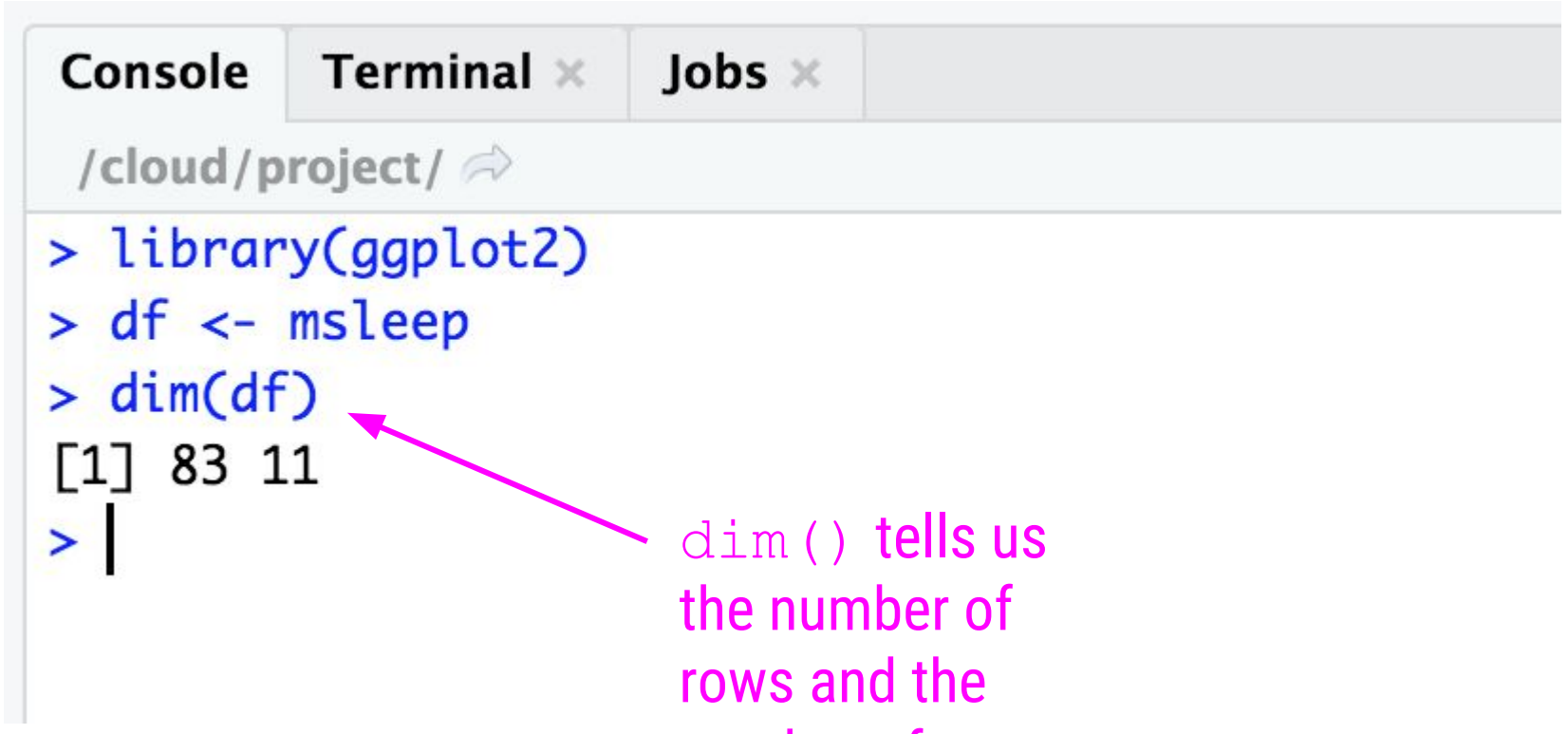
common name

```r
## install and load package
install.packages("ggplot2")
library(ggplot2)

## assign to object `df`
df <- msleep
```

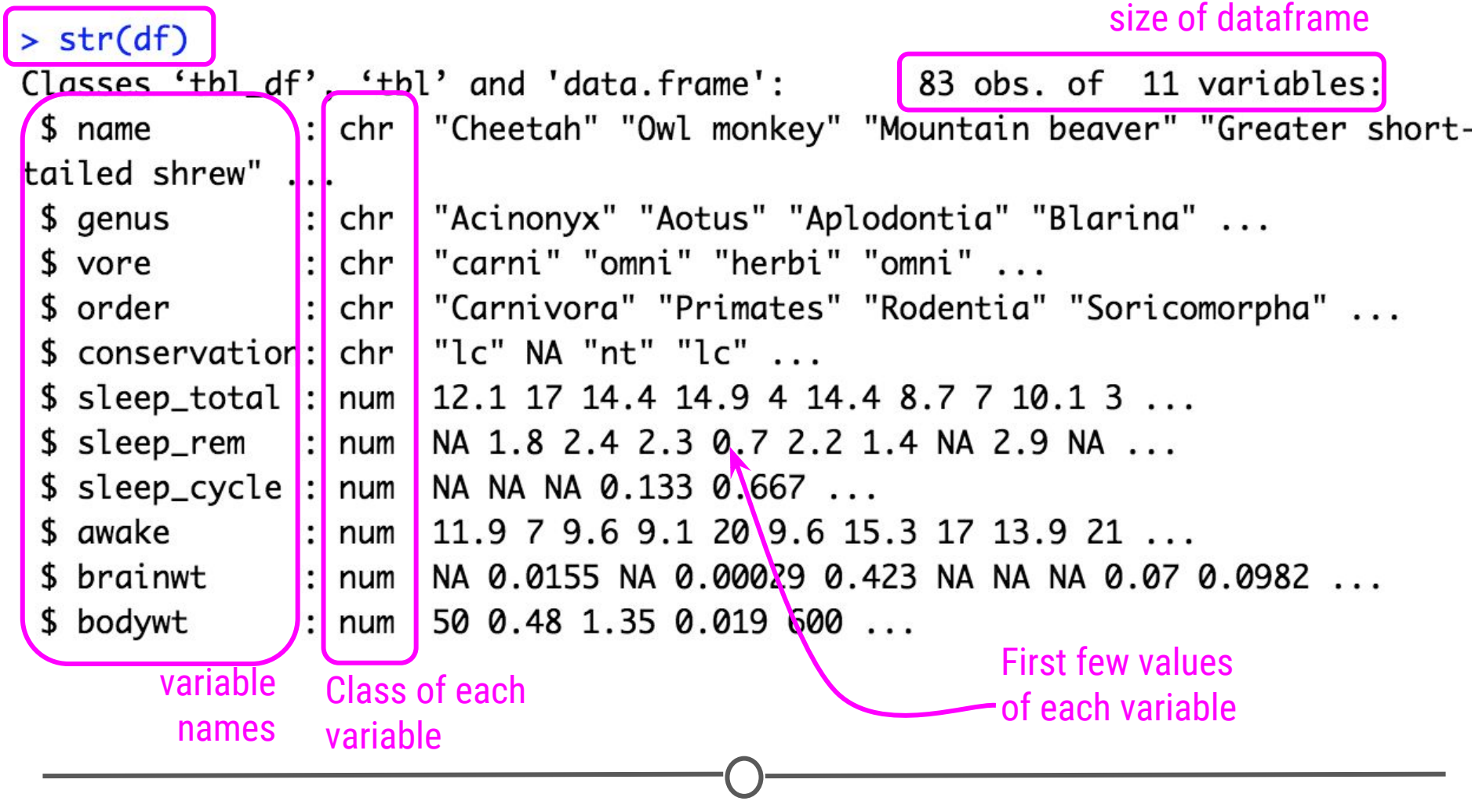


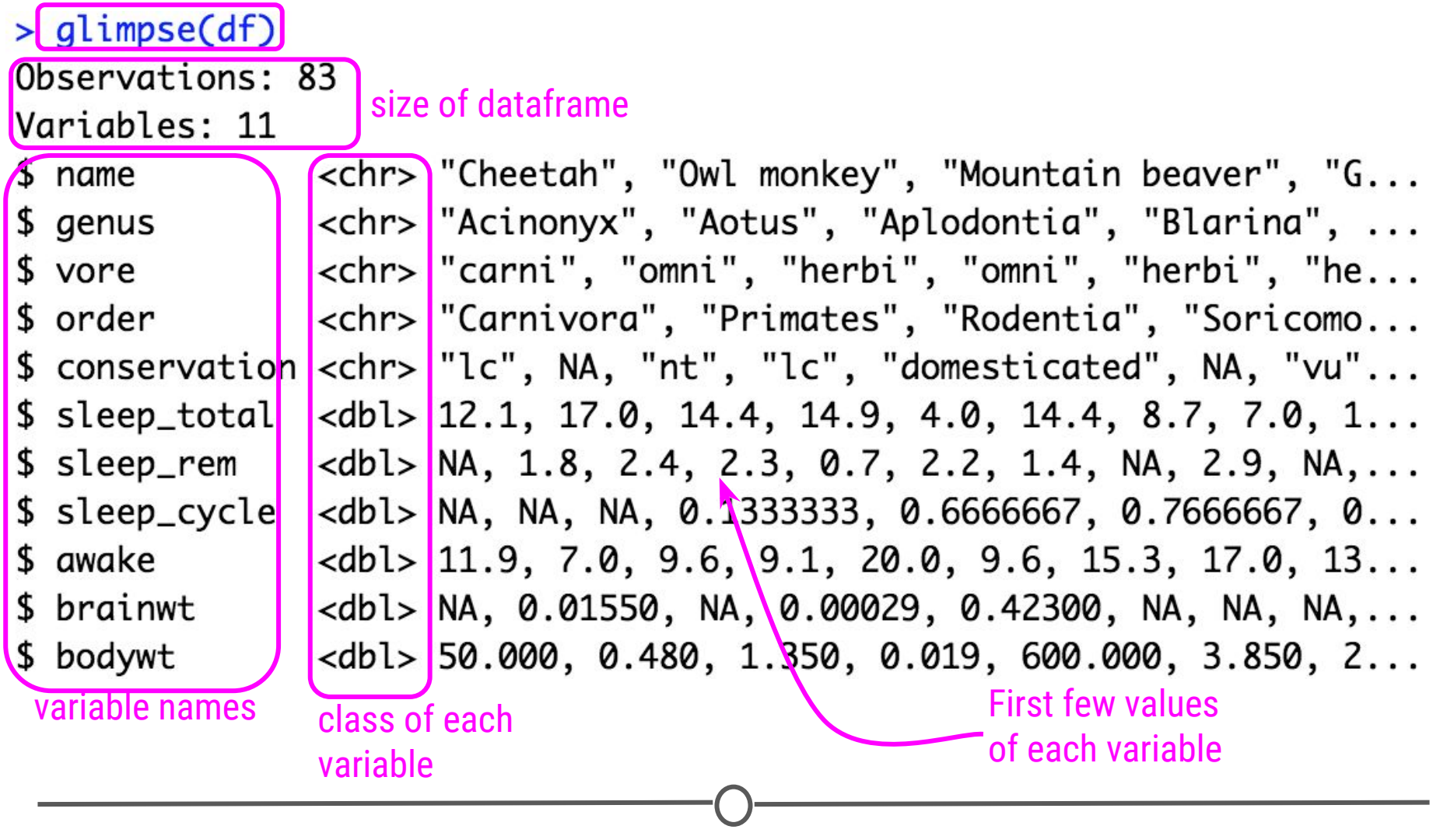
Size

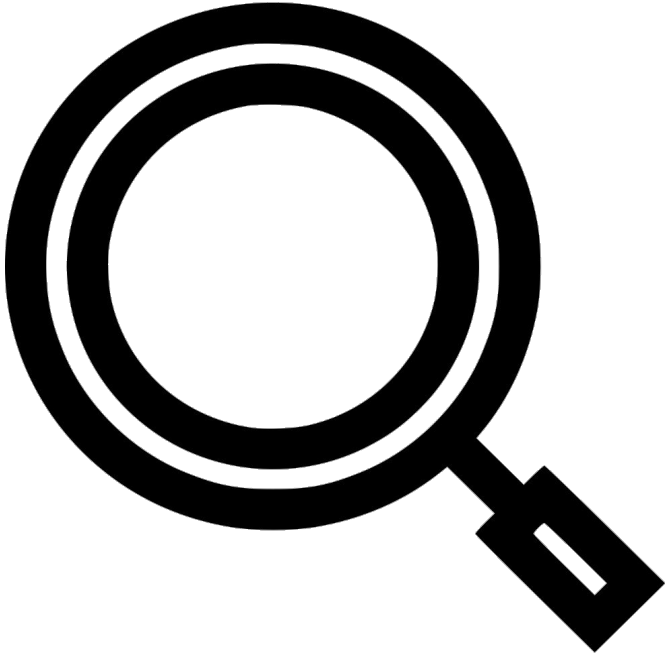

Console
Terminal
Jobs
/cloud/project/
> library(ggplot2)
> df <- msleep
>
Environment
History
Connections
Import Dataset
List
Global Environment
Data
df
83 obs. of 11 variables
Files
Plots
Packages
Help
Viewer
New Folder
Upload
Delete
Rename
More
Cloud
project
Name
Size
Modified
..
Environment tab will tell you
the size of your data frame


Console
Terminal
Jobs
/cloud/project/
> library(ggplot2)
> df <- msleep
> dim(df)
[1] 83 11
>
dim() tells us the number of rows and the number of columns

```
> colnames(df)
 [1] "name"         "genus"        "vore"         "order"
 [5] "conservation" "sleep_total"  "sleep_rem"    "sleep_cycle"
 [9] "awake"        "brainwt"      "bodywt"
```


> str(df)
size of dataframe
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 ...
variable names
Class of each variable
First few values of each variable


> glimpse(df)
Observations: 83
Variables: 11
size of dataframe
$ name <chr> "Cheetah", "Owl monkey", "Mountain beaver", "G...
$ genus <chr> "Acinonyx", "Aotus", "Aplodontia", "Blarina", ...
$ vore <chr> "carni", "omni", "herbi", "omni", "herbi", "he...
$ order <chr> "Carnivora", "Primates", "Rodentia", "Soricomo...
$ conservation <chr> "lc", NA, "nt", "lc", "domesticated", NA, "vu"...
$ sleep_total <dbl> 12.1, 17.0, 14.4, 14.9, 4.0, 14.4, 8.7, 7.0, 1...
$ sleep_rem <dbl> NA, 1.8, 2.4, 2.3, 0.7, 2.2, 1.4, NA, 2.9, NA,...
$ sleep_cycle <dbl> NA, NA, NA, 0.1333333, 0.6666667, 0.7666667, 0...
$ awake <dbl> 11.9, 7.0, 9.6, 9.1, 20.0, 9.6, 15.3, 17.0, 13...
$ brainwt <dbl> NA, 0.01550, NA, 0.00029, 0.42300, NA, NA, NA,...
$ bodywt <dbl> 50.000, 0.480, 1.350, 0.019, 600.000, 3.850, 2...
variable names
class of each variable
First few values of each variable



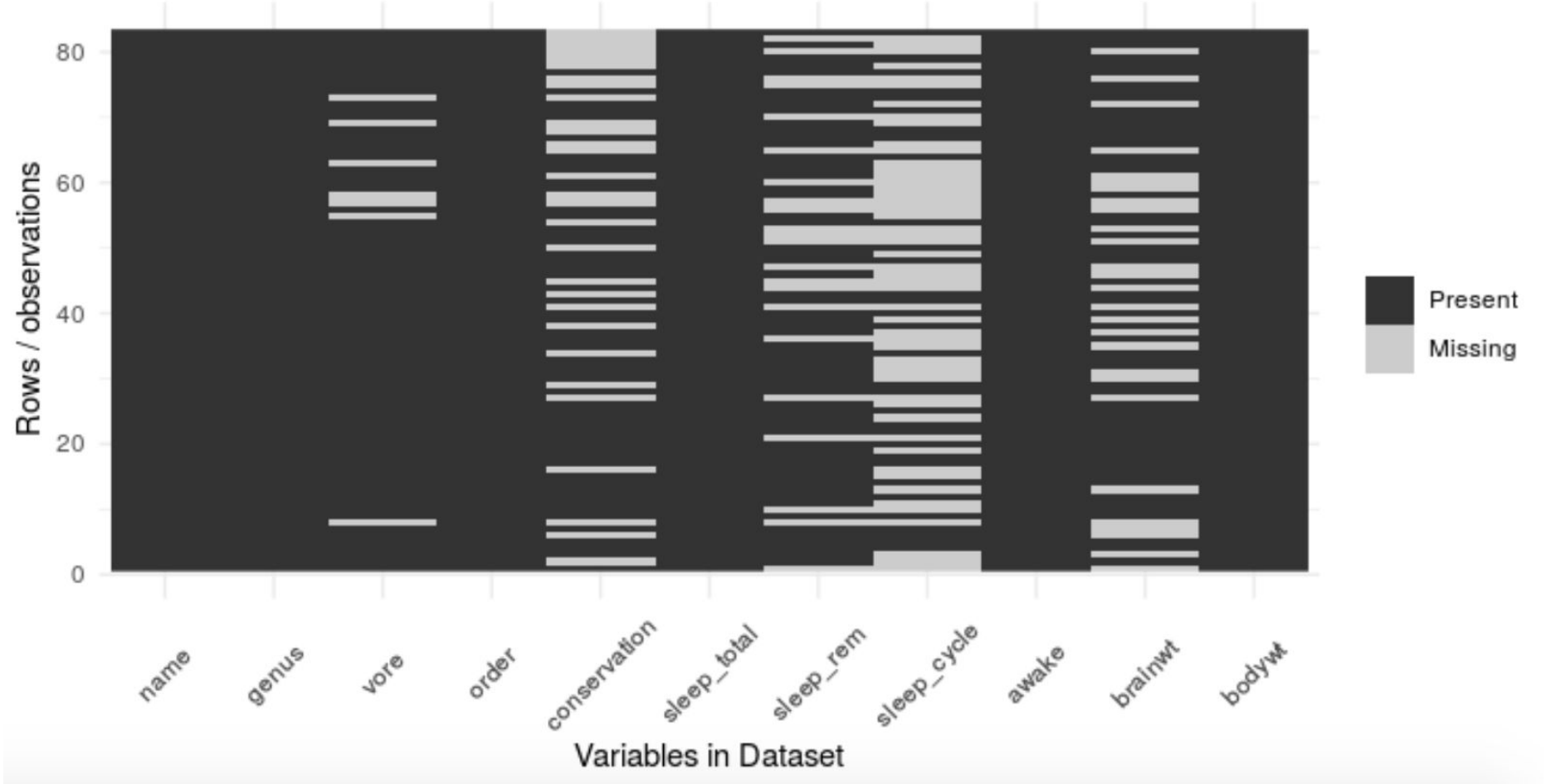
Missingness

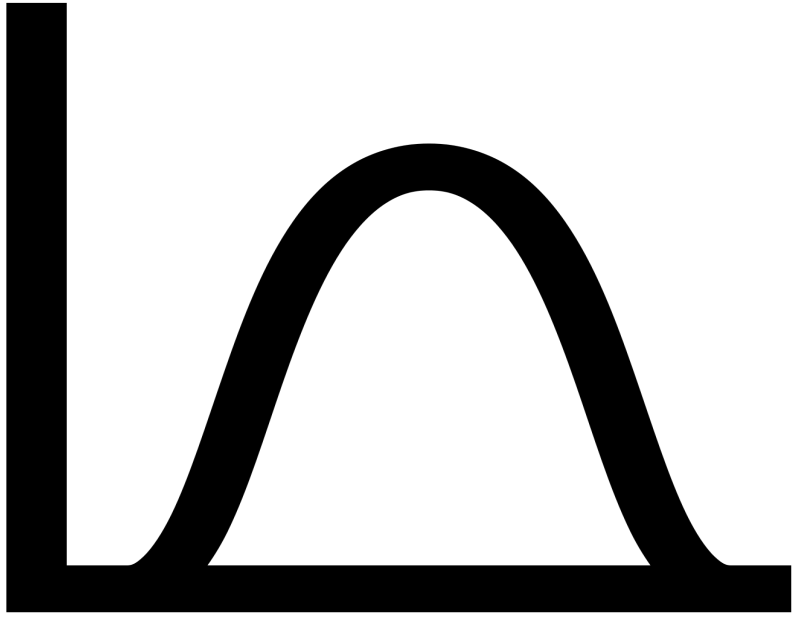
```
> ## calculate how many NAs there are in each variable
> sapply(df, function(x) sum(is.na(x)))
       name         genus          vore         order conservation  sleep_total
          0             0             7             0            29            0
 sleep_rem   sleep_cycle         awake       brainwt       bodywt
         22            51             0            27            0
>
> ## calculate the proportion of missingness
> ## for each variable
> sapply(df, function(x) sum(is.na(x)))/nrow(df)
       name         genus          vore         order conservation  sleep_total
 0.00000000    0.00000000    0.08433735    0.00000000    0.34939759    0.00000000
 sleep_rem   sleep_cycle         awake       brainwt       bodywt
 0.26506024    0.61445783    0.00000000    0.32530120    0.00000000
```

27 observations of `brainwt` are missing

That's 32.5% of the observations in the dataset

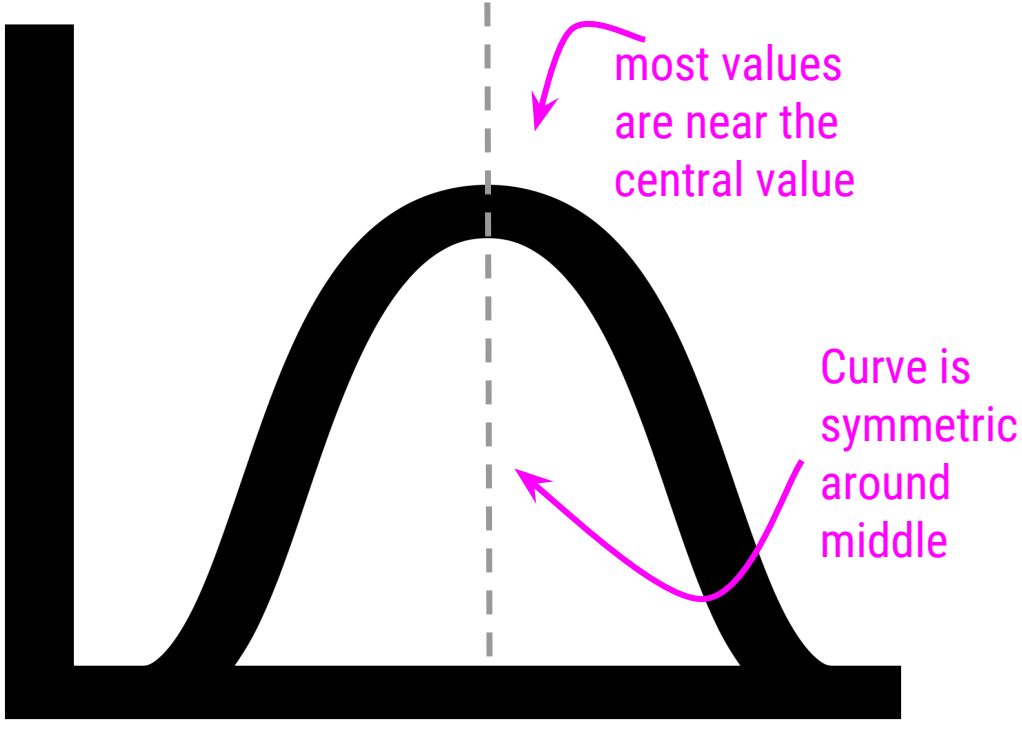
```r
## install and load devtools
install.packages("devtools")
library(devtools)

## install neato package
devtools::install_github("njtierney/neato")
library(neato)

## visualize missingness
ggplot_missing(df)
```


Rows / observations
80
60
40
20
0
name
genus
vore
order
conservation
sleep_total
sleep_rem
sleep_cycle
awake
brainwt
bodywt
Variables in Dataset
Present
Missing

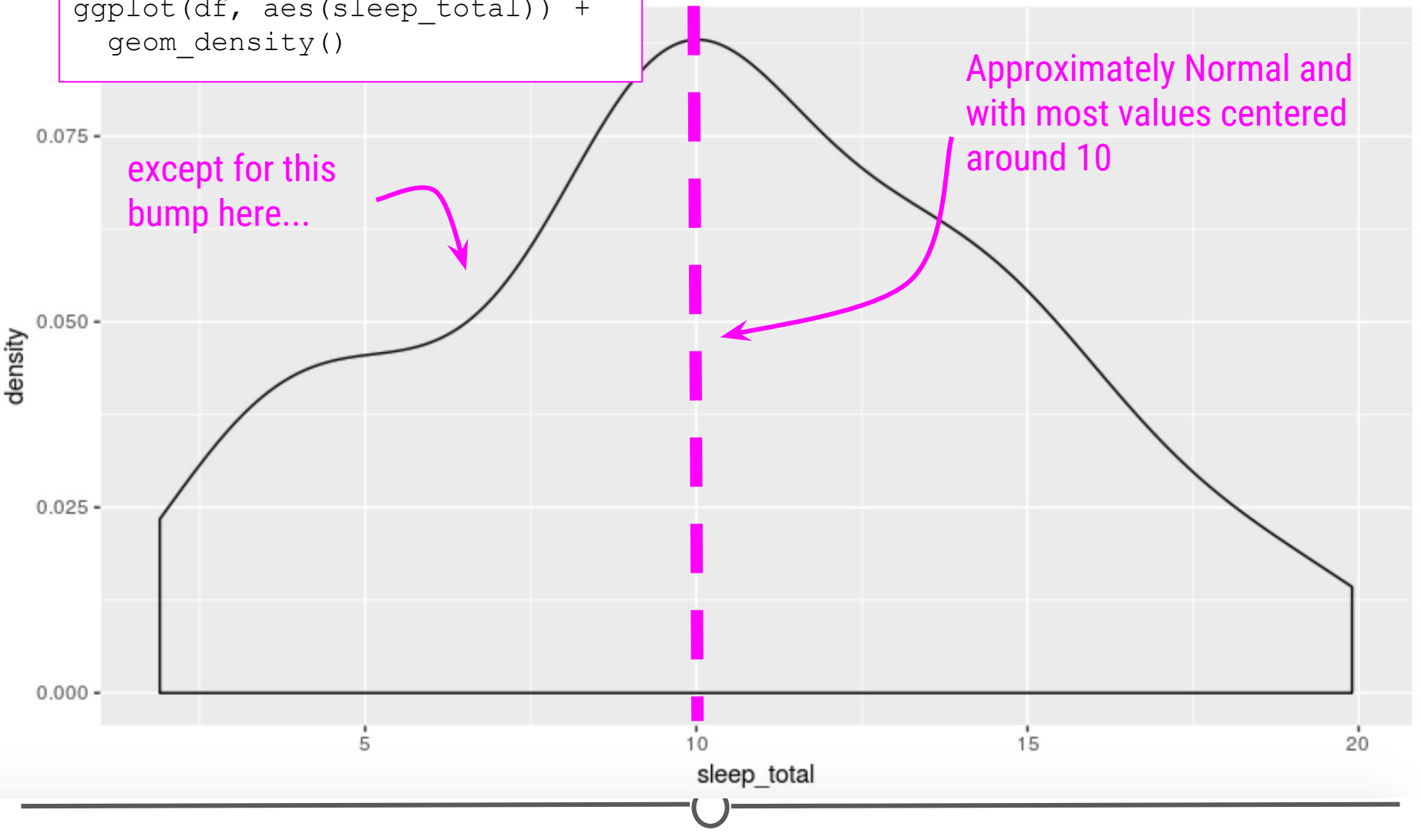


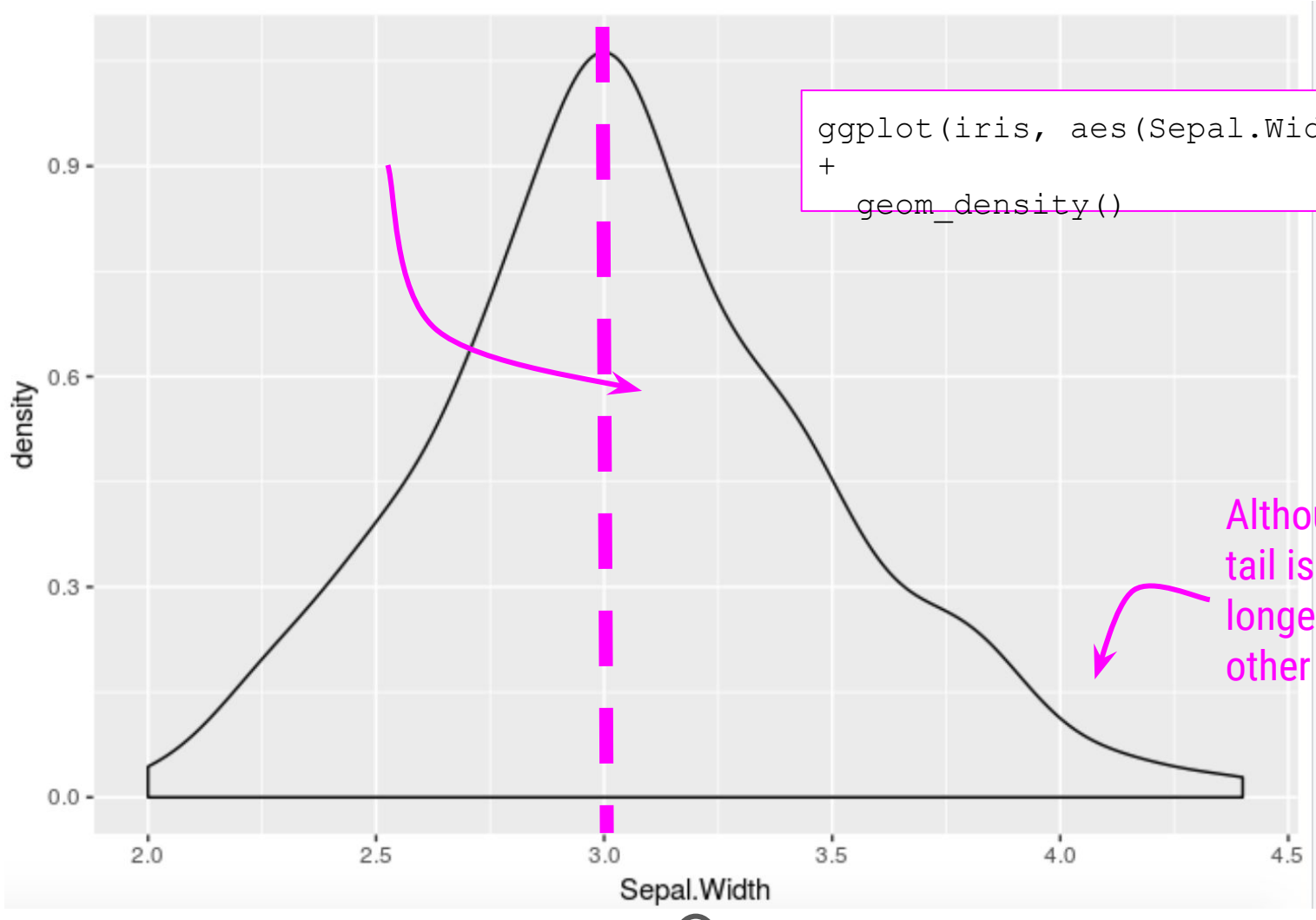
Shape

A Normal Distribution

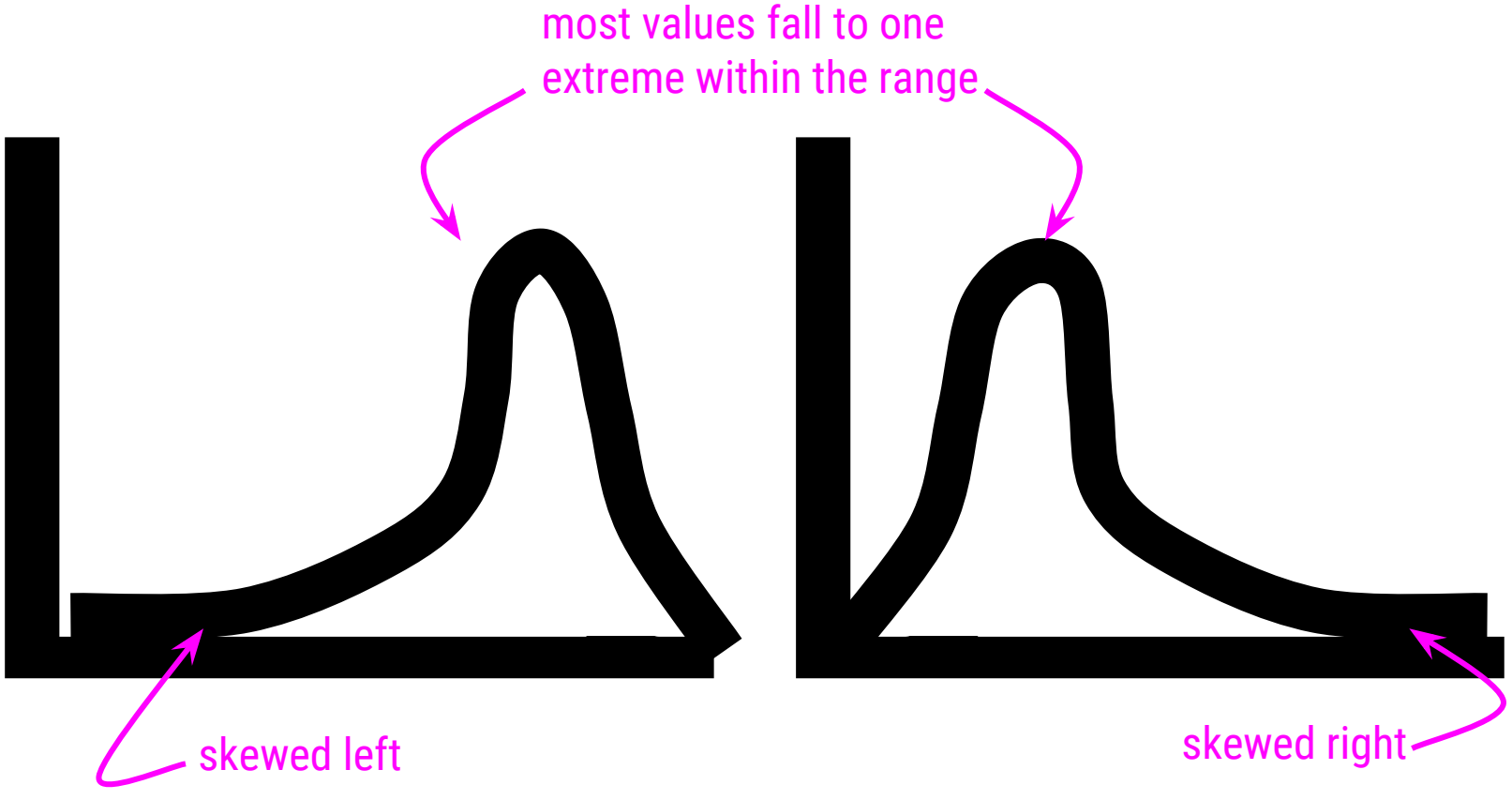


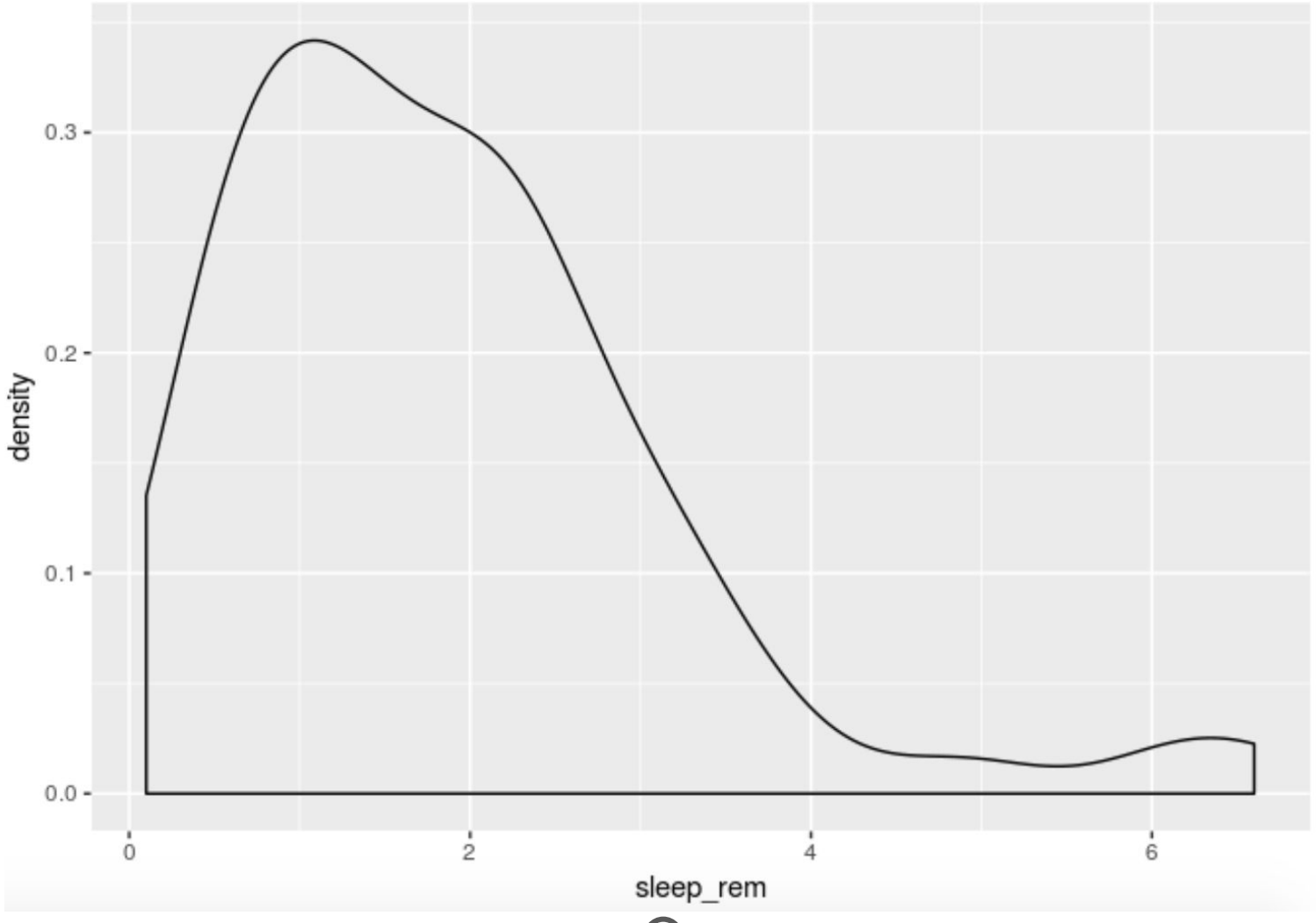
```
ggplot(df, aes(sleep_total)) +
  geom_density()
```



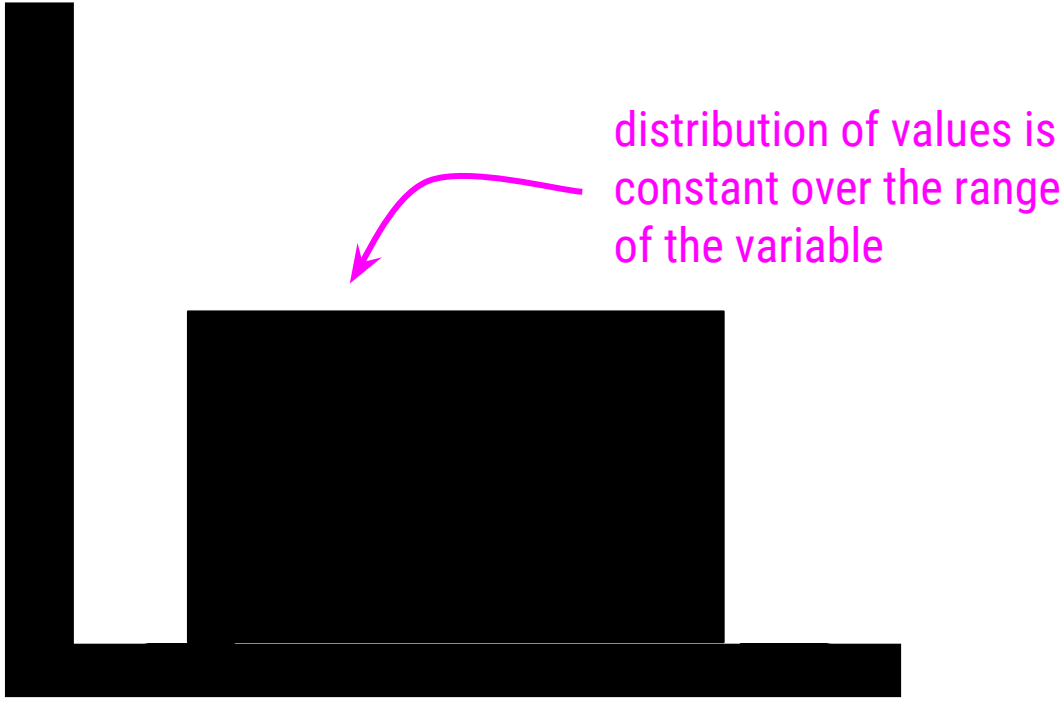

ggplot(iris, aes(Sepal.Width))
+
geom_density()
Although this tail is slightly longer than the other tail
density
0.0
0.3
0.6
0.9
2.0
2.5
3.0
3.5
4.0
4.5
Sepal.Width

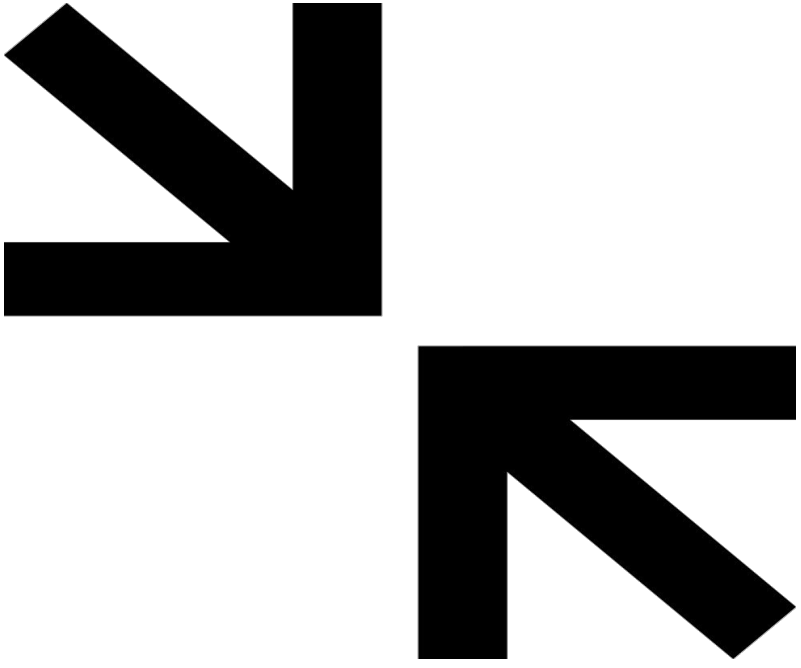
A Skewed Distribution




0.3
0.2
0.1
0.0
density
0
2
4
6
sleep_rem

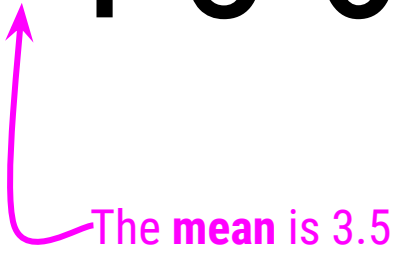
A Uniform Distribution





Central Tendency

1 2 3 4 5 6



Calculating the mean:

1. Sum all values
   1 + 2 + 3 + 4 + 5 +6 = 21
2. Divide sum by the number of observations (6)
   21/6 = **3.5**

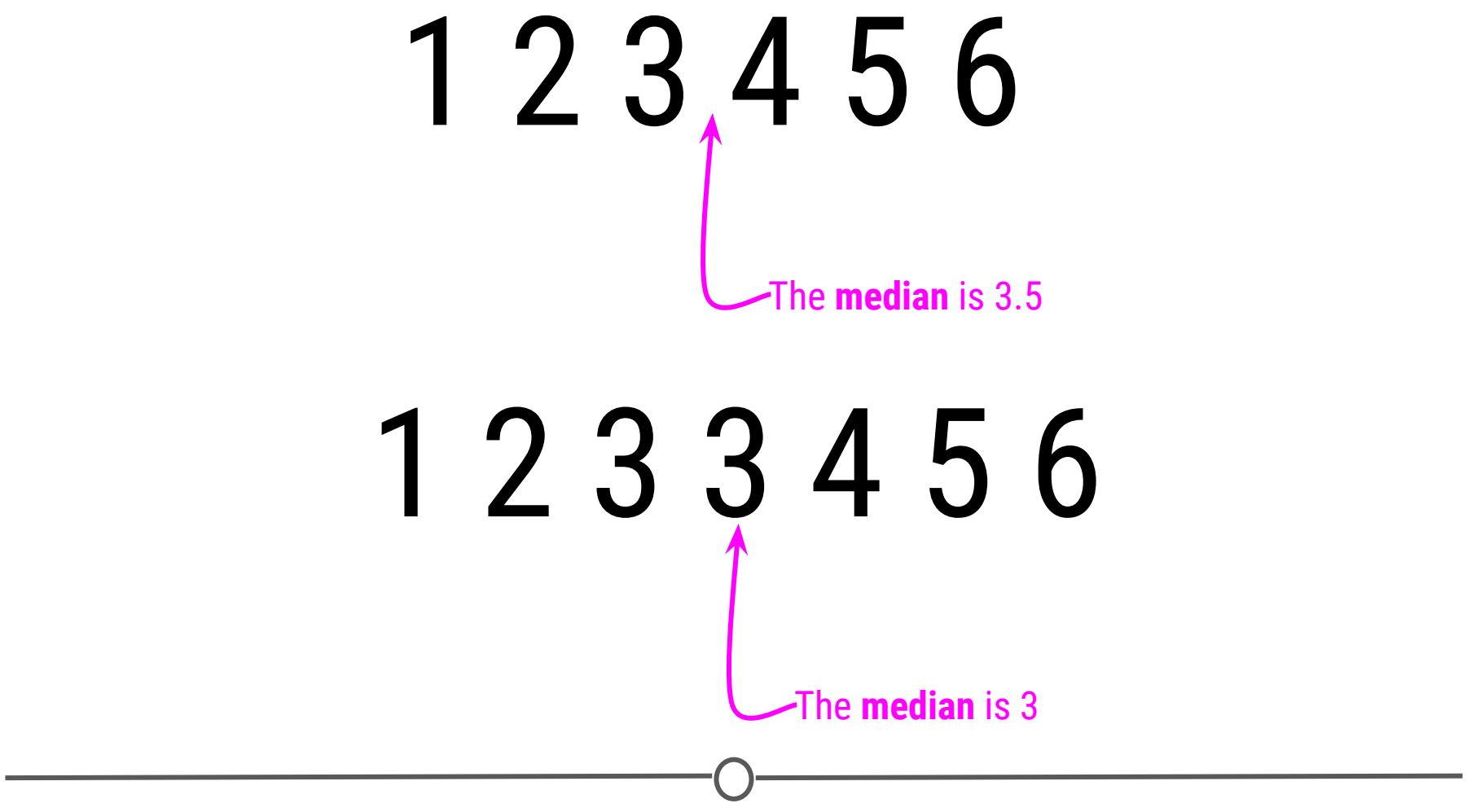
1 2 3 3 4 5 6

Calculating the mean:

1. Sum all values
   1 + 2 + 3 + 3 + 4 + 5 +6 = 24
2. Divide sum by the number of observations (6)
   24/7 = **3.43**

```
> ## this will return NA
> mean(df$sleep_cycle)
[1] NA
>
> ## have to tell R to ignore the NAs
> mean(df$sleep_cycle, na.rm=TRUE)
[1] 0.4395833
```

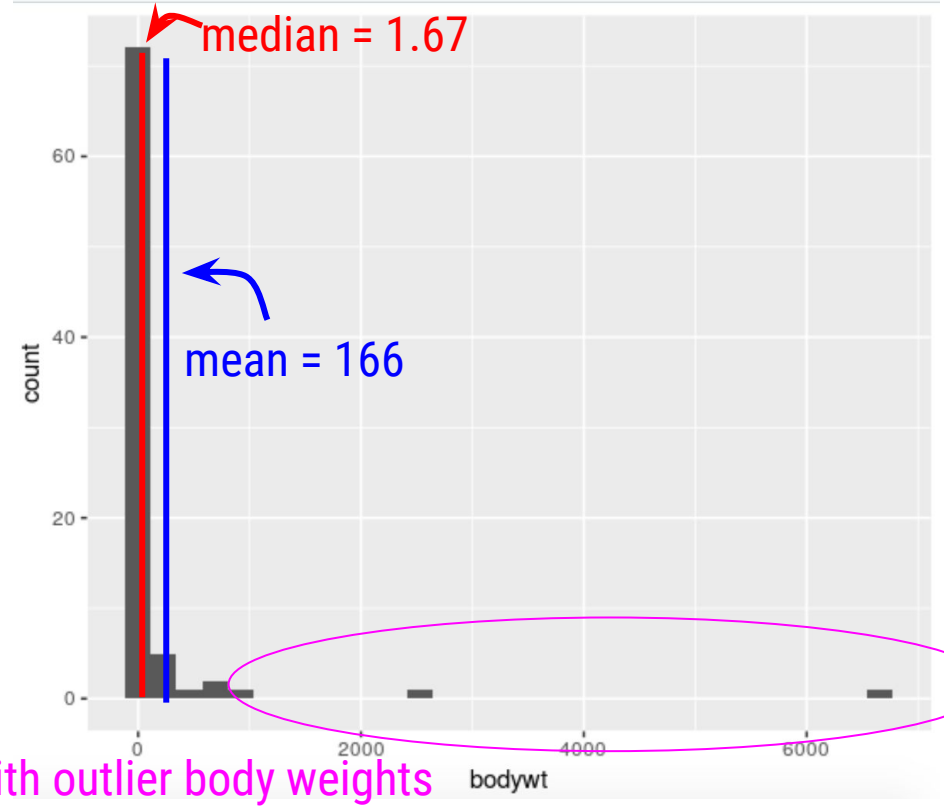
1 2 3 4 5 6


1 2 3 4 5 6
The median is 3.5
1 2 3 3 4 5 6
The median is 3

```
> ## calculate the median
> median(df$sleep_cycle, na.rm=TRUE)
[1] 0.3333333
```

```
ggplot(df,
aes(bodywt)) +
  geom_histogram()
```

```
> mean(df$bodywt)
[1] 166.1363
> median(df$bodywt)
[1] 1.67
```



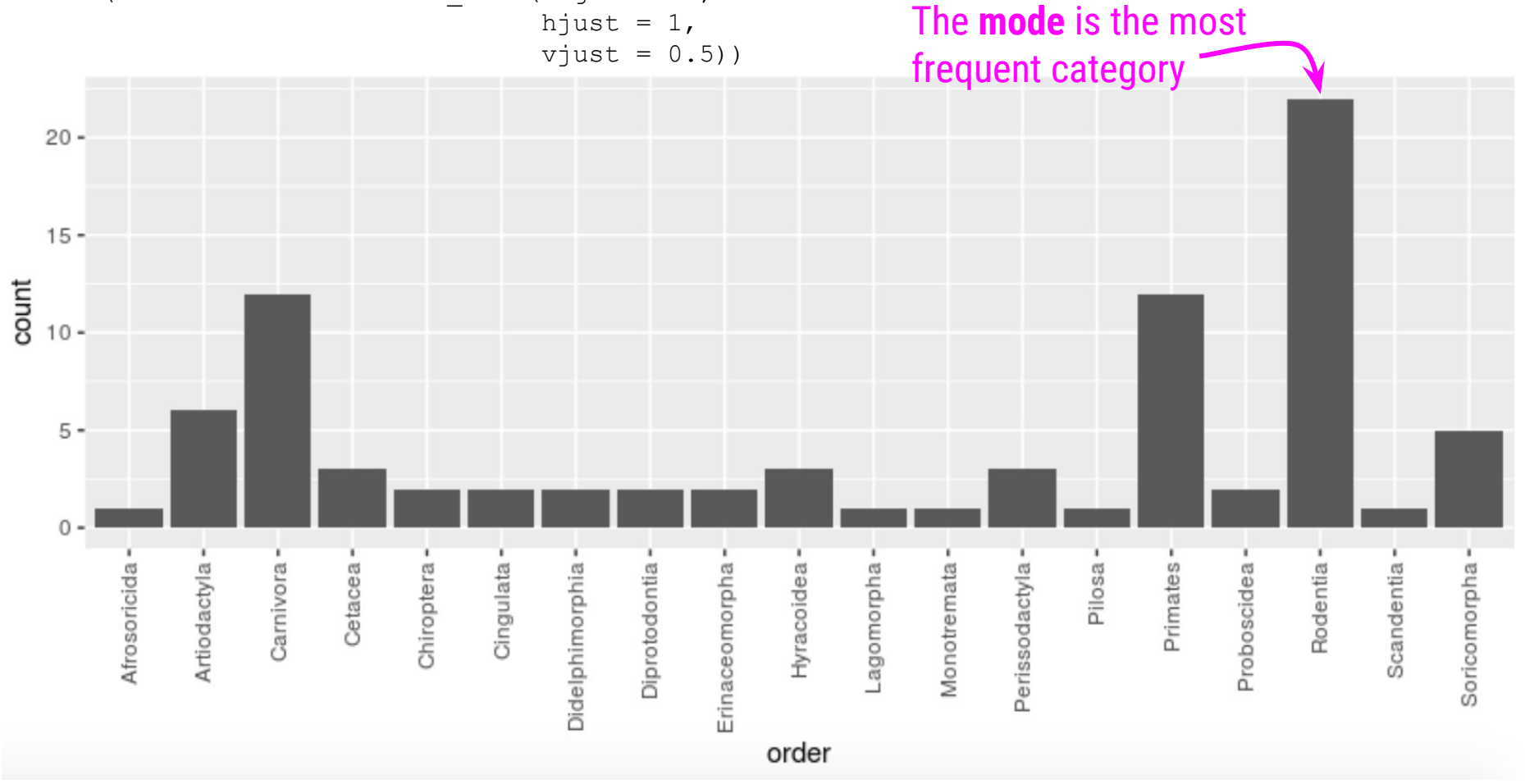
Mammals with outlier body weights lead to an increase in the mean

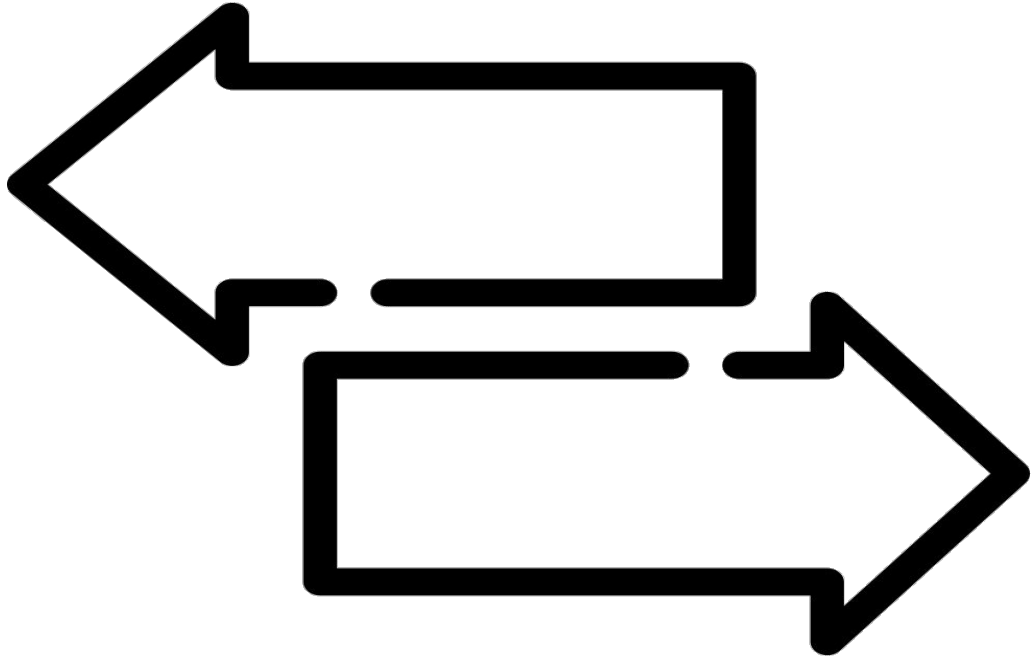
```
> a <- c(0, 10, 10, 3, 5, 10, 10)
> which.max(tabulate(a))
[1] 10
```

```
> table(df$order)

   Afrosoricida     Artiodactyla        Carnivora          Cetacea       Chiroptera
              1                6               12                3                2
      Cingulata Didelphimorphia   Diprotodontia Erinaceomorpha       Hyracoidea
              2                2               2                2                3
     Lagomorpha     Monotremata  Perissodactyla           Pilosa         Primates
              1                1               3                1               12
    Proboscidea         Rodentia     Scandentia    Soricomorpha
              2               22               1                5
```

The **mode** is the most frequent category

```
ggplot(df, aes(order)) +
  geom_bar() +
  theme(axis.text.x = element_text(angle = 90,
                                   hjust = 1,
                                   vjust = 0.5))
```



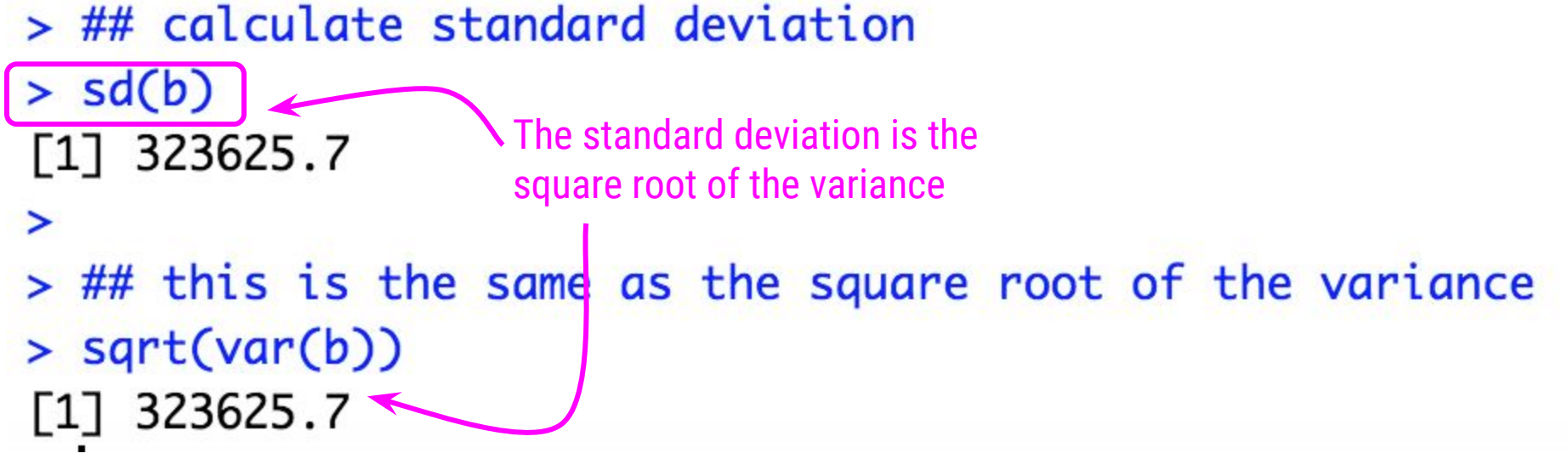


Variability

```
> ## variance of a vector where all values are the same
> a <- c(29, 29, 29, 29)
> var(a)
[1] 0
>
> ## variance of a vector with one very different value
> b <-  c(29, 29, 29, 29, 723678)
> var(b)
[1] 104733575040
```

variance is zero when every value is the same

Large value leads to increased variance


> ## calculate standard deviation
> sd(b)
[1] 323625.7
>
> ## this is the same as the square root of the variance
> sqrt(var(b))
[1] 323625.7
The standard deviation is the
square root of the variance

```
> skim(df)
Skim summary statistics
 n obs: 83
 n variables: 11

── Variable type:character ──────────────────────────────────────────────────────
     variable missing complete  n min max empty n_unique
 conservation      29       54 83   2  12     0        6
        genus       0       83 83   3  13     0       77
         name       0       83 83   3  30     0       83
        order       0       83 83   6  15     0       19
         vore       7       76 83   4   7     0        4

── Variable type:numeric ────────────────────────────────────────────────────────
    variable missing complete  n   mean     sd      p0     p25    p50   p75    p100     hist
       awake       0       83 83  13.57   4.45 4.1     10.25   13.9  16.15   22.1
      bodywt       0       83 83 166.14 786.84 0.005    0.17    1.67 41.75 6654
     brainwt      27       56 83   0.28   0.98 0.00014 0.0029  0.012 0.13    5.71
 sleep_cycle      51       32 83   0.44   0.36 0.12    0.18    0.33  0.58    1.5
   sleep_rem      22       61 83   1.88   1.3  0.1     0.9     1.5   2.4     6.6
 sleep_total       0       83 83  10.43   4.45 1.9     7.85    10.1  13.75   19.9
```

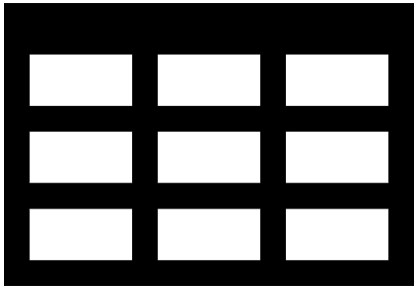
missingness

variability

shape

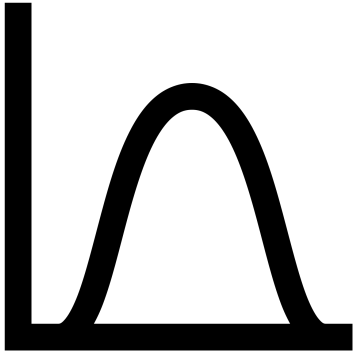
central tendency

Descriptive Analysis

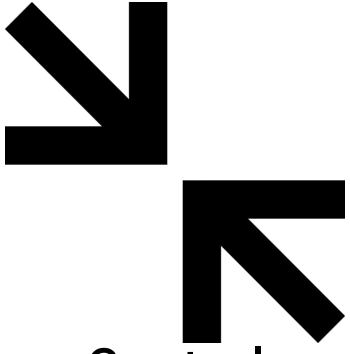


Size

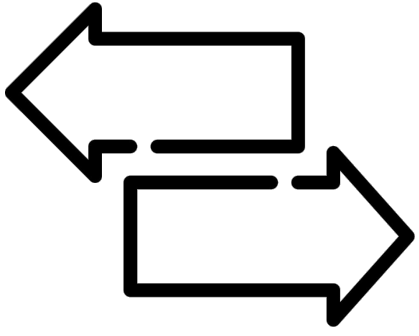
Missingness



Shape



Central Tendency



Variability