

Quarto, data frames, and basic plots



Overview

Quick review from last class

- R as a calculator, objects, data types
- Functions

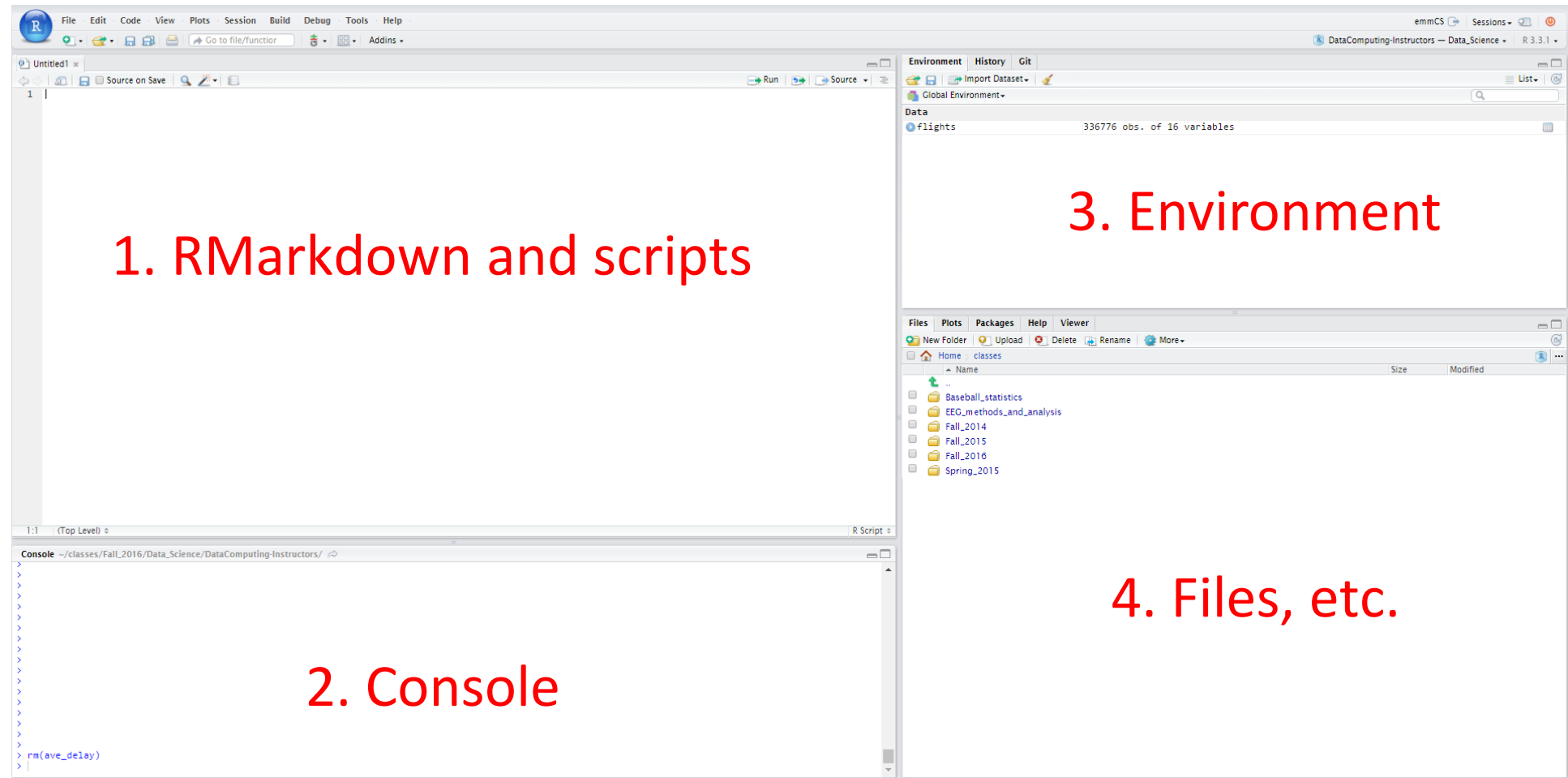
More R

- Vectors and packages
- Quarto
- Data Frames

If there is time:

- Plots and statistics of categorical data
- Plots and statistics of quantitative data

Please open up RStudio



Review: R Basics

Arithmetic:

```
> 2 + 2
```

```
> 7 * 5
```

Assignment of values to ***objects***:

```
> a <- 4
```

```
> b <- 7
```

```
> z <- a + b
```

```
> z
```

```
[1] 11
```

Review: Character strings and Booleans

```
> a <- 7
```

```
> s <- "s is a terrible name for an object"
```

```
> b <- TRUE
```

```
> class(a)
```

```
[1] numeric
```

```
> class(s)
```

```
[1] character
```

Review: Functions

Functions use parenthesis: functionName(x)

```
> sqrt(49)
```

```
> tolower("DATA is AWESOME!")
```

To get help

```
> ? sqrt
```

One can add comments to your code

```
> sqrt(49)  # this takes the square root of 49
```

Vectors

Vectors are ordered sequences of numbers or letters

The `c()` function is used to create vectors

```
> v <- c(5, 232, 5, 543)
```

```
> s <- c("statistics", "data", "science", "fun")
```

One can access elements of a vector using square brackets `[]`

```
> s[4]      # what will the answer be?
```

Vectors continued

One can also apply functions to vectors

```
> z <- 2:10
```

```
> sqrt(z)
```

```
> min(z)
```

```
> which.min(z)
```

We can also add and subtract vectors of the same length

```
> v1 <- c(2, 4, 8)
```

```
> v2 <- c(1, 2, 3)
```

```
> v1 + v2
```


R packages

R packages

Packages add additional functionality to R

We will use many additional packages in this class

- ggplot2, dplyr, tidyr, etc.

There is a class specific package (SDS111) I wrote that you can use to download homework and other files

- All class materials are also on GitHub: <https://github.com/emeyers/SDS111>



SDS111 package

If you are using R and Rstudio on your own computer, instructions to install the SDS111 package are at: <https://github.com/emeyers/SDS111>

- Sonam and I can help you with this too during office hours

If you are using RStudio on the YCRC cluster, the SDS111 package is already installed

To use the functions in the package you can use

```
library(SDS111)
```

```
download_homework(1)
```

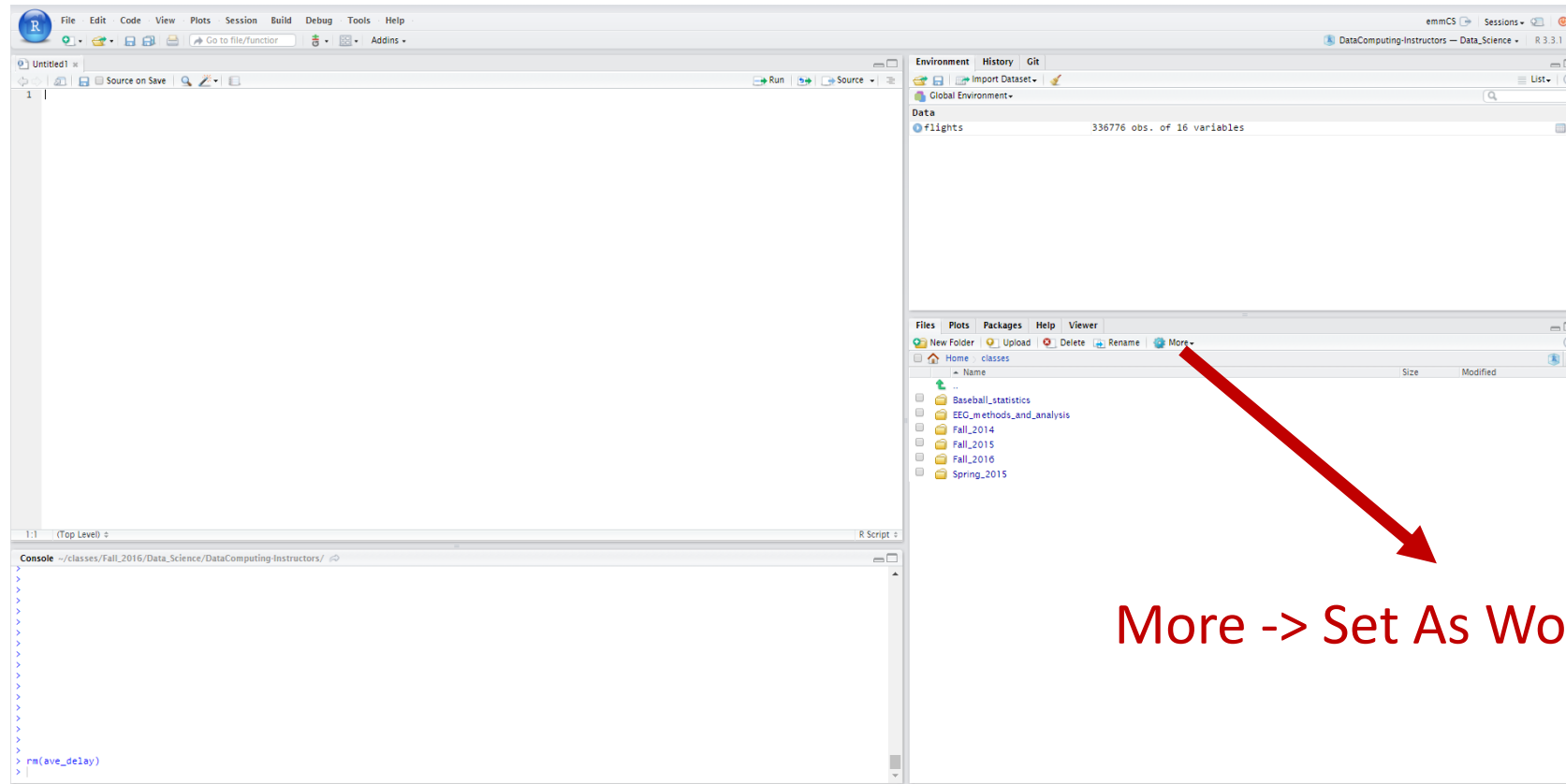
Downloading class 2 code

We can use the SDS111 package to get code for today's class by typing the following commands at the console:

```
> library(SDS111)
```

```
> download_class_code(2)
```

Setting your working directory



More -> Set As Working Directory

1. In the files tab, navigate to the directory that contains the .qmd file
2. Click More -> Set As Working Directory

Questions?



Quarto

Quarto

Quarto (.qmd files) allow you to embed written descriptions, R code and the output of that code into a nice looking document

Creates a way to do reproducible research!



Quarto

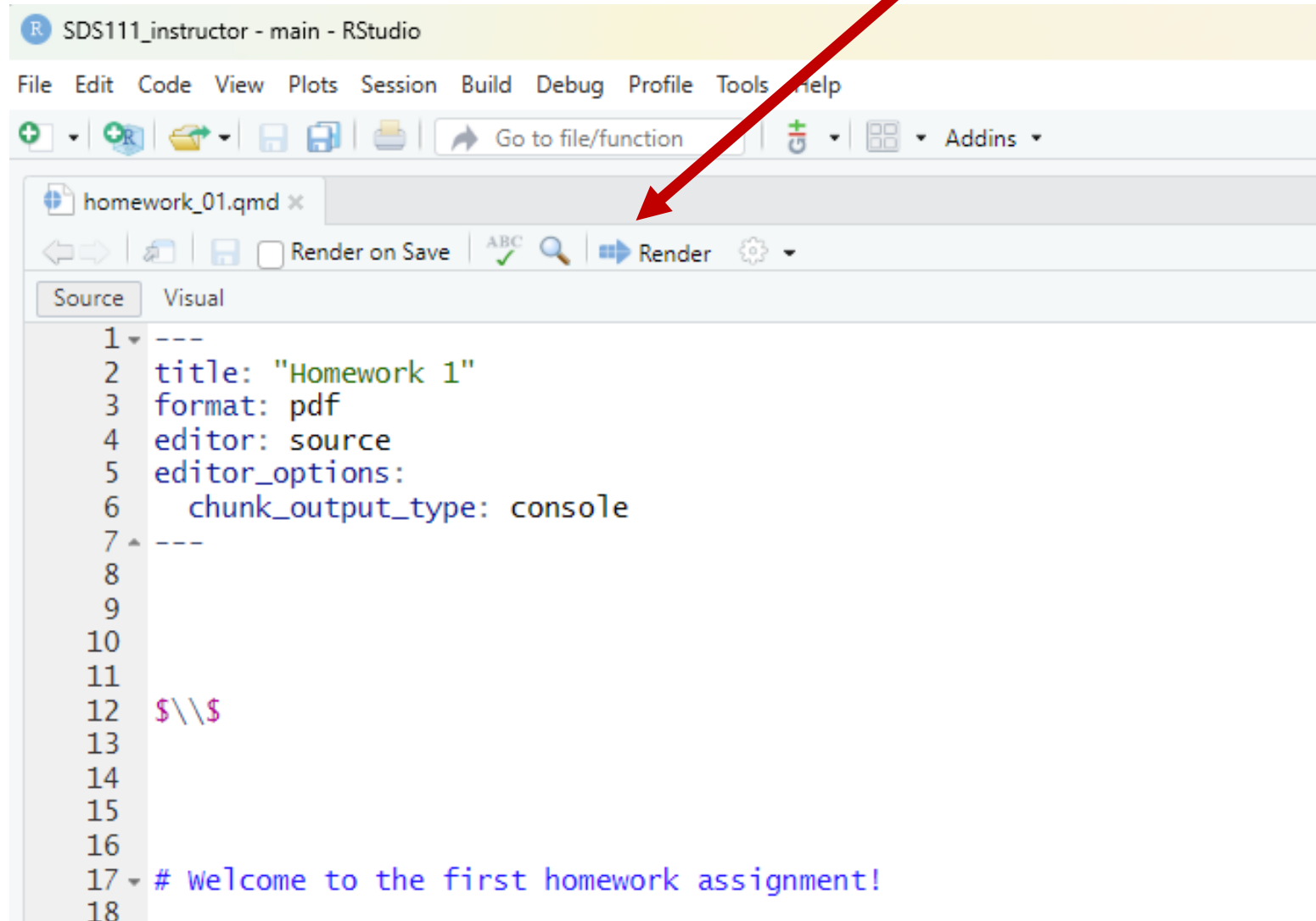
Everything in R chunks is executed as code:

```
```${r}  
 # this is a comment
 # the following code will be executed
 2 + 3
```
```

Everything outside R chunks appears as text

Render to a pdf

Turn in a pdf or html document
with your solutions to Canvas



Quarto

Note: When you render a Quarto document, your Quarto document **does not have access to variables in the global environment**, but instead have their own environment.

Why is this a good thing???

Formatting in Quarto

We can add formatting to text outside the code chunks

Examples:

`## Level 2 header`

`**bold**`

``

To repeat: avoid hard to debug code!

Only change a few lines at a time and then render your document to make sure everything is working!

If your document isn't rendering:

- **For code chunks:** use the `# symbol` to comment out code until you can find the line of code that is giving the error message
- **Outside of code chunk:** cut out part of the document until it renders and then paste it back

Announcement: Homework 1

Due Monday July 7th at 11pm

- I recommend getting started early on this!

To download the homework please do the following:

```
> library(SDS111)
```

```
> download_homework(1)
```

From the file panel, open the homework and try knitting it

Announcement: Homework 1

Instructions for how to submit homework on Gradescope are on Canvas

- Please mark all pages that answers correspond to on Gradescope!

Be sure to also "show your work" by printing out any values you report

- Although don't print out hundreds of access pages of numbers


Ask/answer questions on Ed Discussions, but don't give away the solutions!

Questions?



Data frames

Data frames contain structured data

|  | age | body_type | diet | drinks | drugs | education |
|---|-----|----------------|-------------------|----------|-----------|-----------------------------------|
| 1 | 22 | a little extra | strictly anything | socially | never | working on college/university |
| 2 | 35 | average | mostly other | often | sometimes | working on space camp |
| 3 | 38 | thin | anything | socially | NA | graduated from masters program |
| 4 | 23 | thin | vegetarian | socially | NA | working on college/university |
| 5 | 29 | athletic | NA | socially | never | graduated from college/university |
| 6 | 29 | average | mostly anything | socially | NA | graduated from college/university |

Back to R: Data frames

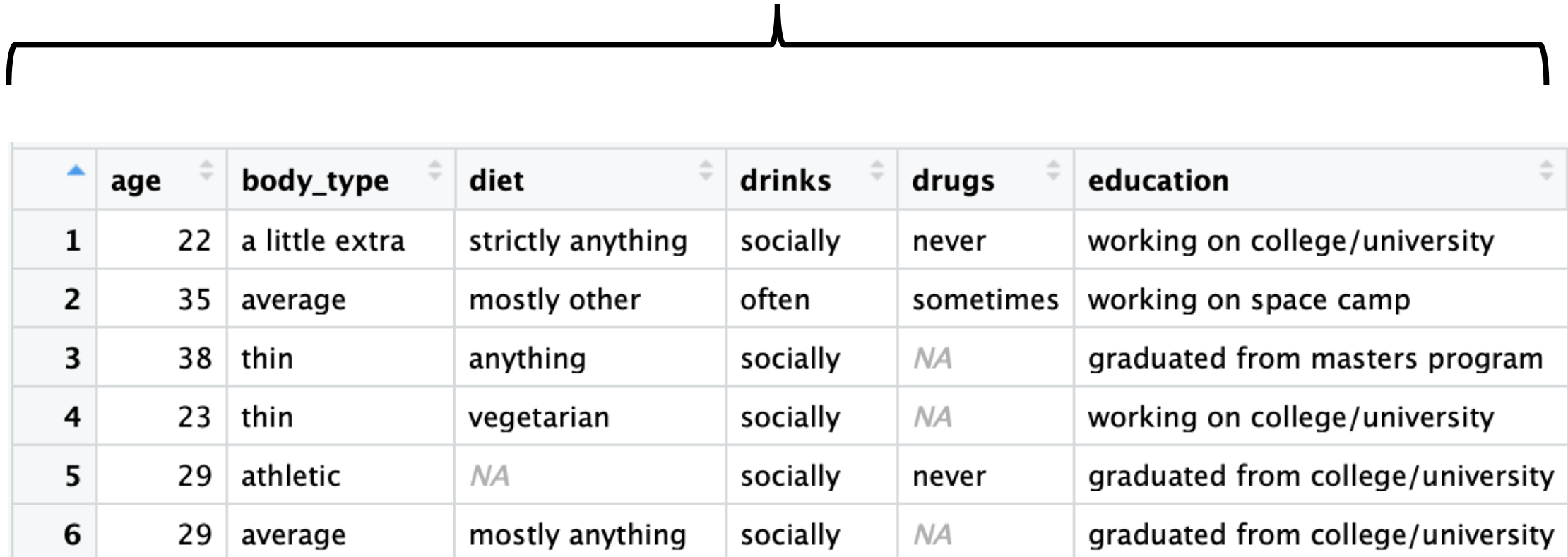
Data frames contain structured data

```
> library(SDS111)
> download_data("profiles_revised.csv") # only needs to be run once
> profiles <- read.csv("profiles_revised.csv")
> View(profiles) # the View() function only works in R Studio!
```

| | age | body_type | diet | drinks | drugs | education |
|---|-----|----------------|-------------------|----------|-----------|-----------------------------------|
| 1 | 22 | a little extra | strictly anything | socially | never | working on college/university |
| 2 | 35 | average | mostly other | often | sometimes | working on space camp |
| 3 | 38 | thin | anything | socially | NA | graduated from masters program |
| 4 | 23 | thin | vegetarian | socially | NA | working on college/university |
| 5 | 29 | athletic | NA | socially | never | graduated from college/university |
| 6 | 29 | average | mostly anything | socially | NA | graduated from college/university |

Data Frames

Variables



| | age | body_type | diet | drinks | drugs | education |
|---|-----|----------------|-------------------|----------|-----------|-----------------------------------|
| 1 | 22 | a little extra | strictly anything | socially | never | working on college/university |
| 2 | 35 | average | mostly other | often | sometimes | working on space camp |
| 3 | 38 | thin | anything | socially | NA | graduated from masters program |
| 4 | 23 | thin | vegetarian | socially | NA | working on college/university |
| 5 | 29 | athletic | NA | socially | never | graduated from college/university |
| 6 | 29 | average | mostly anything | socially | NA | graduated from college/university |

Cases

An Example Dataset

Quantitative Variable

Categorical Variable

Cases
(observational units)

| | age | body_type | diet | drinks | drugs | education |
|---|-----|----------------|-------------------|----------|-----------|-----------------------------------|
| 1 | 22 | a little extra | strictly anything | socially | never | working on college/university |
| 2 | 35 | average | mostly other | often | sometimes | working on space camp |
| 3 | 38 | thin | anything | socially | NA | graduated from masters program |
| 4 | 23 | thin | vegetarian | socially | NA | working on college/university |
| 5 | 29 | athletic | NA | socially | never | graduated from college/university |
| 6 | 29 | average | mostly anything | socially | NA | graduated from college/university |

Data frames

We can extract the columns of a data frame as vector objects using the \$ symbol

```
> the_ages <- profiles$age
```

Can you get the sum of the ages of users in this data set?

```
> sum(the_ages)
```

Let's try it in RStudio!

Questions?



Categorical data

Categorical variables

What is a categorical variable?

- A: A categorical variable assigns each observation to one of k groups

Which variables in the profiles data frame are categorical?

- Is age a categorical variable?

For categorical variables, we usually want to view:

- How many items are each category OR
- The proportion (or percentage) of items in each category

$$\text{Proportion in a category} = \frac{\text{number in that category}}{\text{total number}}$$

Categorical data

```
# Get information about drinking behavior
```

```
> drinking_vec <- profiles$drinks
```

```
# Create a table showing how often people drink
```

```
> drinks_table <- table(drinking_vec)
```

```
> drinks_table
```

Relative frequency table

We can create a relative frequency table using the function:

```
> prop.table(my_table)
```

Can you create a relative frequency table for the drinking behavior of the people in the okcupid data set?

```
> drinks_table <- table(profiles$drinks)
```

```
> prop.table(drinks_table)
```

Bar plots

(pun intended?)

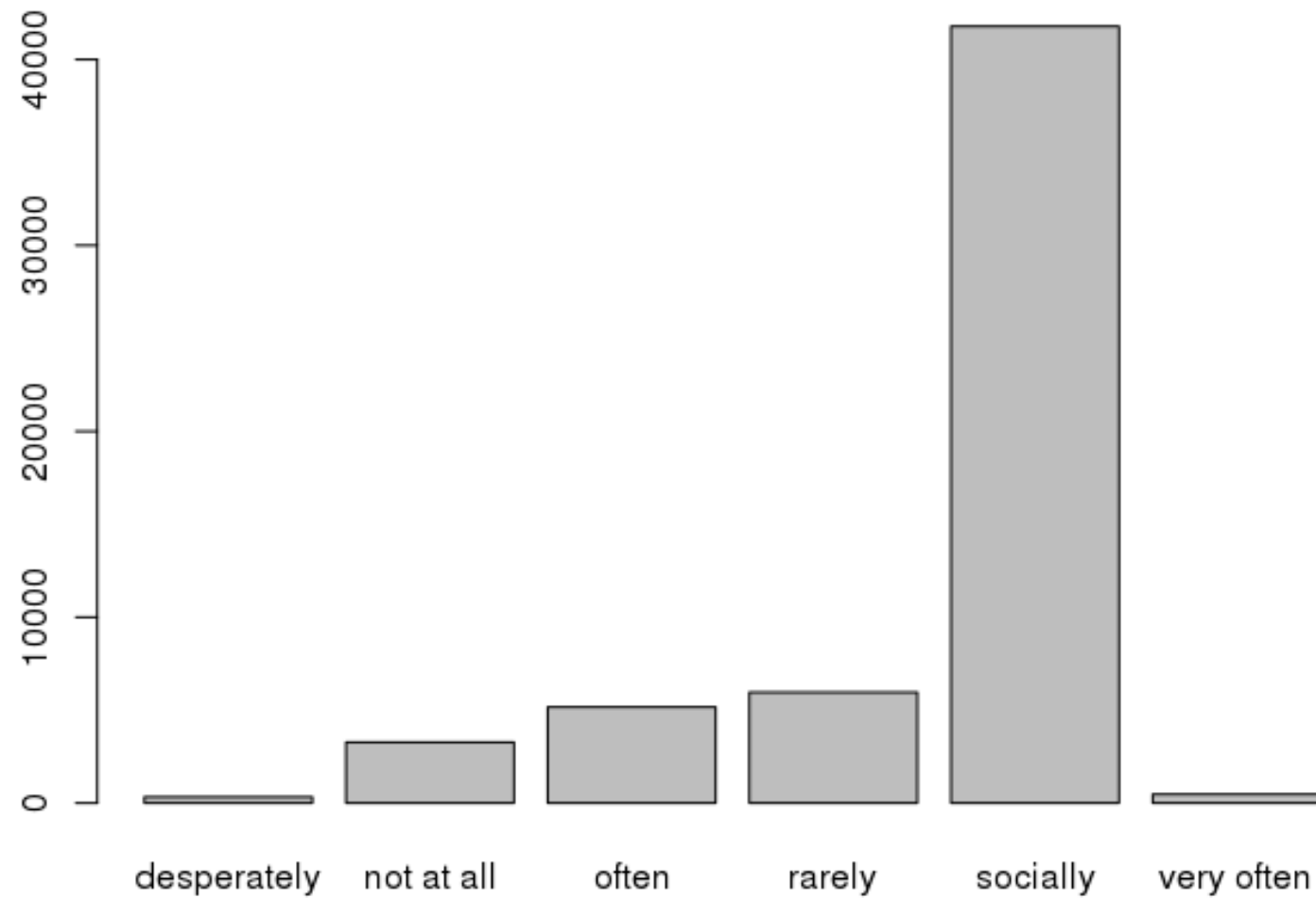
We can plot the number of items in each category using a bar plot

```
> barplot(my_table)
```

Can you create a bar plot for the drinking behavior of the people in the okcupid data set?

```
> drinks_table <- table(profiles$drinks)
```

```
> barplot(drinks_table)
```



What is wrong with this plot?

Details matter!

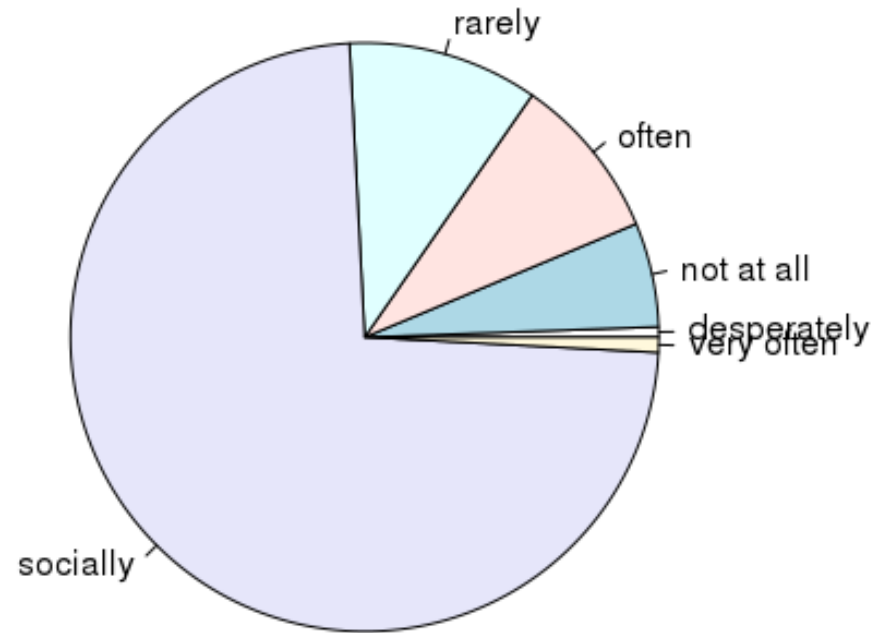
Can you figure out how to label the axes?

```
> barplot(drinks_table,  
          ylab = "Count",  
          xlab = "Type of drinker",  
          main = "Counts of different types of drinkers")
```

Pie charts

We can also use the `pie()` function to create pie charts

```
> pie(drinks_table)
```



Which is best: bar plots or pie charts?

```
> barplot(table(profiles$sex, useNA = "always"))
```

```
> pie(table(profiles$sex, useNA = "always"))
```

Q1: Is one better than the other?

Q2: Can you figure out how to add colors to these plots?

Questions?



Visualizing Quantitative Data

Visualizing quantitative data: histograms

The first few okcupid users' heights

- > `profiles$height[1:5]`
- 75, 70, 68, 76, ...

To create a histogram we create a set of intervals

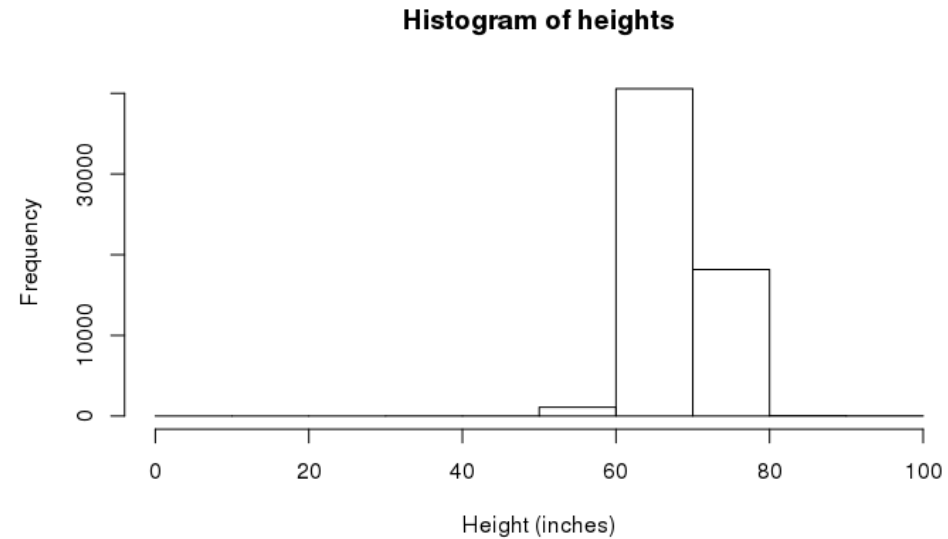
- 60-65, 65-70, 70-75, 75-80

We count the number of points that fall in each interval

We create a bar chart with the counts in each bin

Histograms of heights

| Height (inches) | Frequency Count |
|-----------------|-----------------|
| (0-10] | 6 |
| (10-20] | 0 |
| (20-30] | 1 |
| (30-40] | 13 |
| (40-50] | 9 |
| (50-60] | 1097 |
| (60-70] | 40575 |
| (70-80] | 18164 |
| (80-90] | 50 |
| >90 | 28 |



Visualizing heights

We can create histograms in R using the `hist()` function

Can you create a histogram of heights?

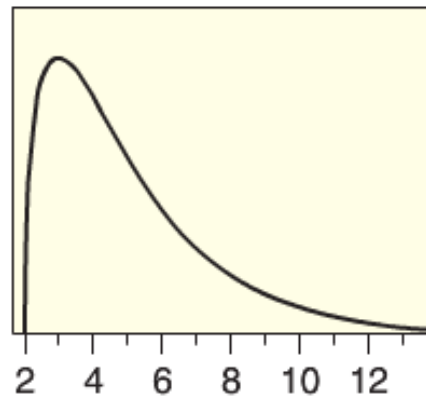
```
> hist(profiles$height)
```

How can you add more bars to the histogram?

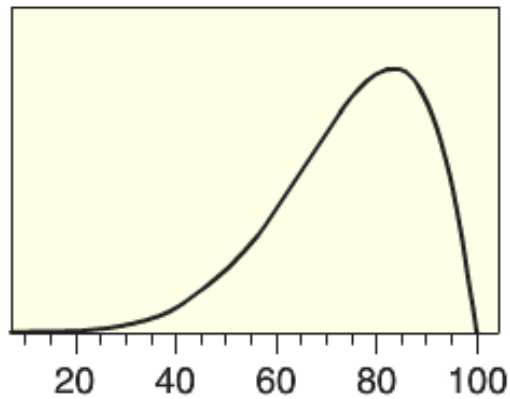
- How can we figure out how to add more bars to a histogram?

```
> hist(profiles$height, breaks = 50)
```

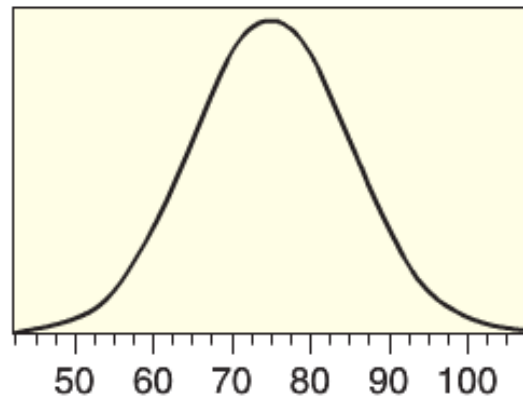
Common shapes for distributions



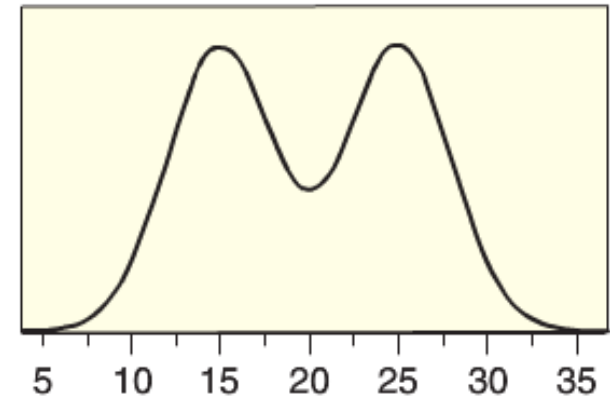
(a) Skewed to the right



(b) Skewed to the left



(c) Symmetric and bell-shaped



(d) Symmetric but not bell-shaped

Statistics for quantitative data

Measure of central tendency: The mean

$$\text{Mean} = \frac{\text{Sum of all data values}}{\text{Number of data values}}$$

$$\text{Mean} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n} = \sum_{i=1}^n \frac{x_i}{n} = \frac{1}{n} \sum_{i=1}^n x_i$$

R: `mean(x)`

R: `mean(x, na.rm = TRUE)`

The median

The **median** of a data set of size n is

- If n is odd: The middle value of the sorted data
- If n is even: The average of the middle two values of the sorted data

The median splits the data in half

```
R: median(v)  
    median(v, na.rm = TRUE)
```


Resistance

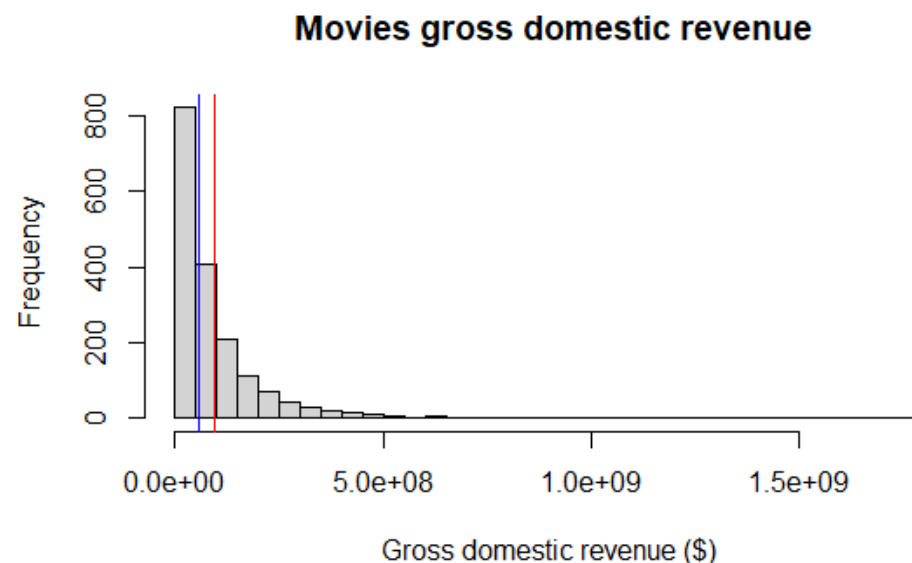
We say that a statistics is **resistant** if it is relatively unaffected by extreme values (outliers).

The median is resistant when the mean is not

Example:

Mean US salary = \$72,641

Median US salary = \$51,939



For next class...

Homework 1 is due on Gradescope by 11pm on Monday July 7th

- Instructions for how to submit homework on Gradescope are on Canvas