R Markdown, data frames and plots

Overview

Back to R basics

R Markdown

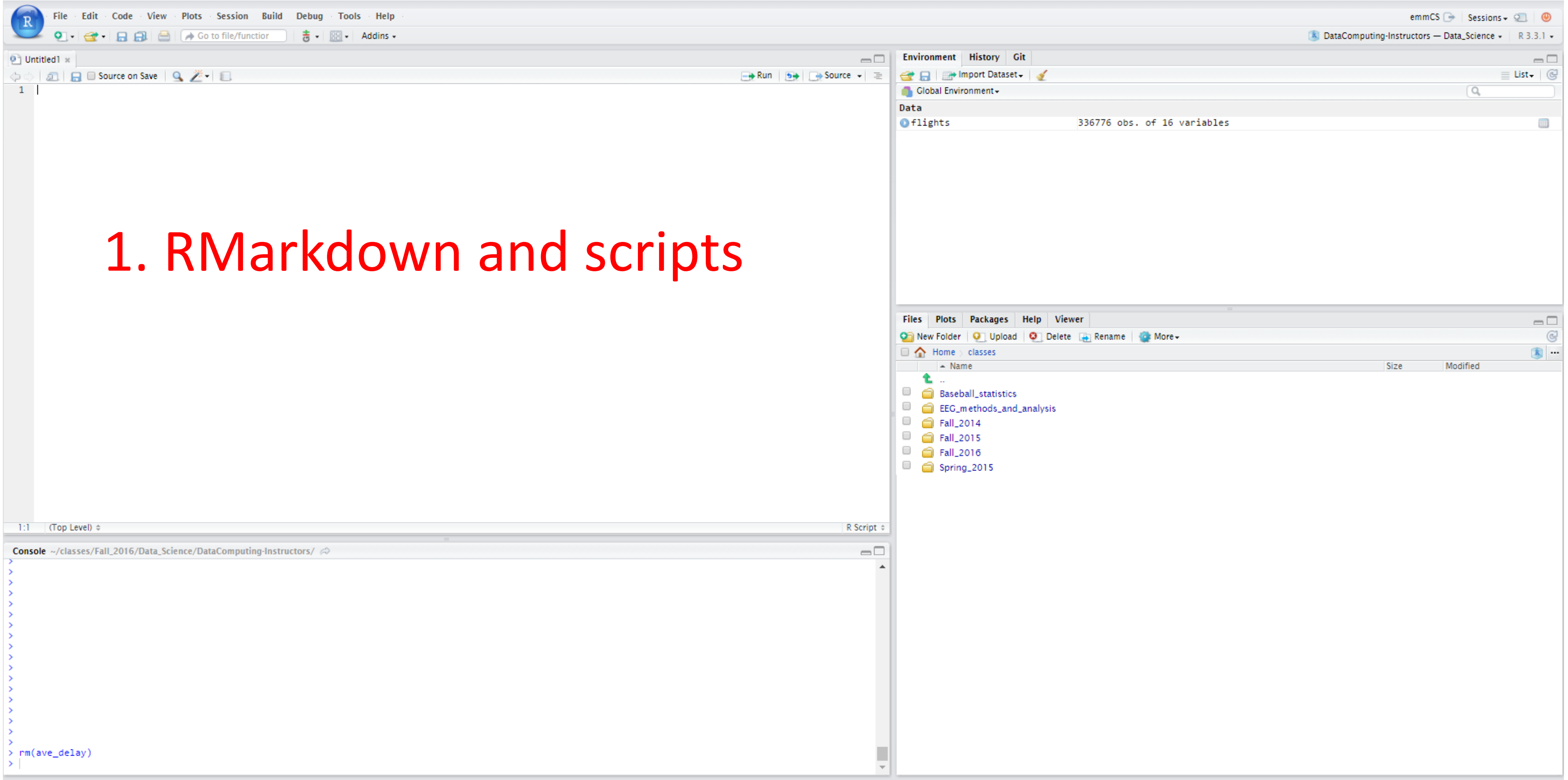
- Formatting
- Code Chunks

More R

- Data frames
- Categorical data: statistics and plots
- Quantitative data: statistics and plots

Any questions about anything?

RStudio layout



Console ~/classes/Fall_2016/Data_Science/DataComputing-Instructors/

```
> rm(ave_delay)
>
```

```
flights        336776 obs. of 16 variables
```

1. RMarkdown and scripts

Create a new script

File -> New File -> R Script

Save the script with a reasonable name, e.g., week1_notes.R

R Basics

Arithmetic:

```
>   2 + 2
>   7 * 5
```

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

```
>  a <- 4
>  b <- 7
>  z  <- a + b
>  z
[1]  11
```

Number journey…

Character strings and booleans

```
> a <- 7
> s <- "s is a terrible name for an object"
> b <- TRUE

> class(a)

> class(s)
```

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?
```

We can get multiple elements from a vector too

```
> s[c(1, 2)]
```

Vectors continued

One can assign a sequence of numbers to a vector

```
> z <- 2:10
> z[3]
```

One can test which elements are greater than a value

```
> z > 3
```

Can add names to vector elements

```
> names(v) <- c("first", "second", "third", "fourth")
```

Vectors continued

One can also apply functions to vectors

```
> z <- 2:10
> sqrt(z)
> mean(z)
```

Questions?

Question

Q: What kind of grades the pirate get in Introduction to Statistics?

Q: Worst joke of the semester?

R packages

Packages add additional functionality to R

We will use many additional packages in this class

- gplyr, ggplot2, tidyr, etc.

There is a class specific package (SDS230) I wrote that you can use to download homework and other files

- All class materials are also on GitHub: https://github.com/emeyers/SDS230

Was everyone able to install the SDS230 package?

Downloading class 2 code

If you have the class SDS230 package, you can get code for today's class by typing the following commands at the console:

```
>  library(SDS230)

> download_class_code(2)
```

R Markdown

R Markdown

R Markdown (.Rmd 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!

R Markdown

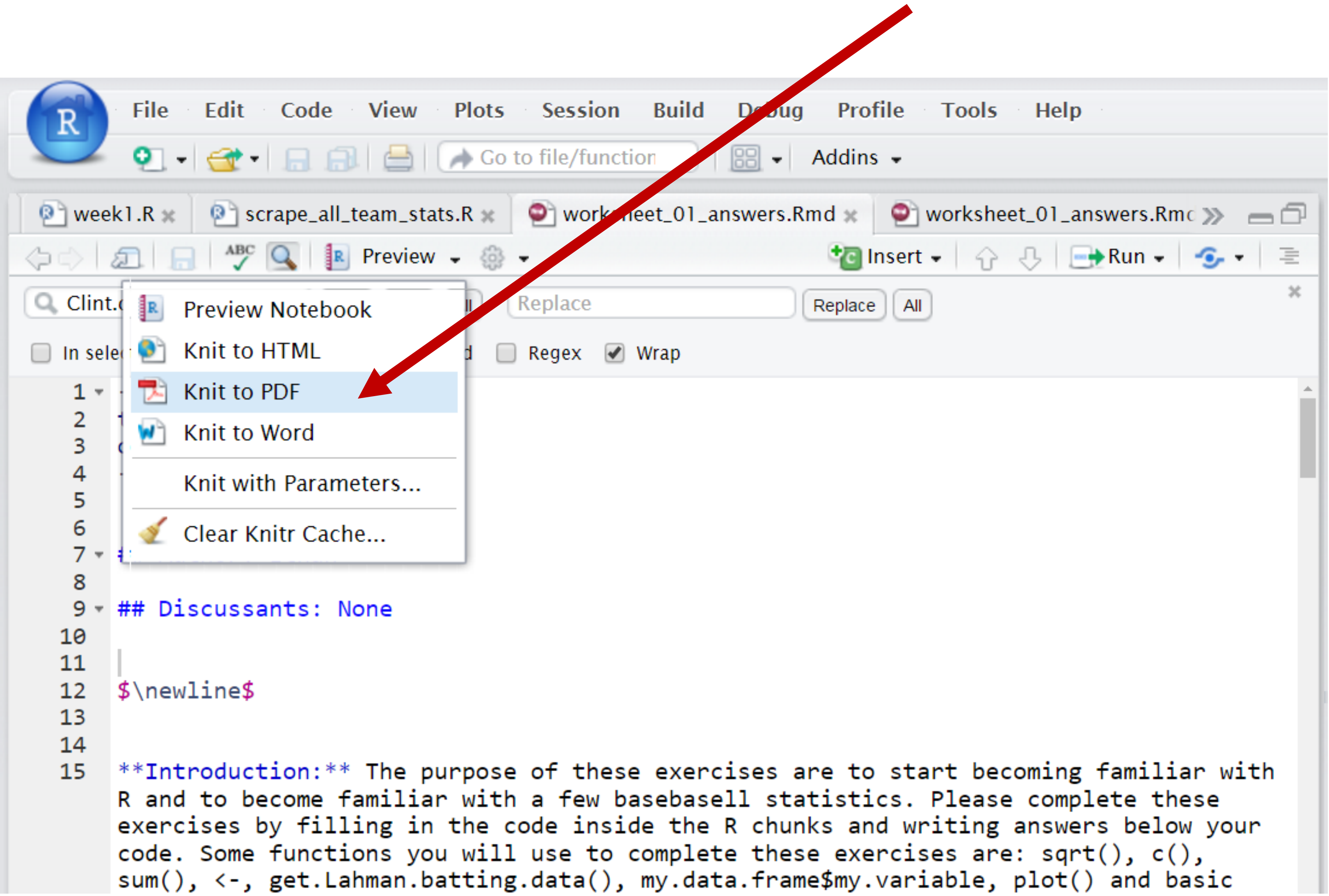
Everything in R chunks is executed as code:

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

Everything outside R chunks appears as text

Knitting to a pdf

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



```
## Discussants: None


$\newline$


**Introduction:** The purpose of these exercises are to start becoming familiar with
R and to become familiar with a few basebasell statistics. Please complete these
exercises by filling in the code inside the R chunks and writing answers below your
code. Some functions you will use to complete these exercises are: sqrt(), c(),
sum(), <-, get.Lahman.batting.data(), my.data.frame$my.variable, plot() and basic
```

R Markdown

Note: When you knit, RMarkdown files **do not have access to variables in the global environment**, but instead have their own environment.

Why is this a good thing???

Formatting in R Markdown

We can add formatting to text outside the code chunks

Examples:

```
## Level 2 header
**bold**
![](https://statistics.yale.edu/sites/default/files/logo2.png)
```

LaTeX in R Markdown

We can also add LaTeX symbols to documents using $\symbol$ syntax

For example, try these:

θ

$\hat{p}$

$\hat{\theta}$

Knit early and knit often to avoid errors!!!

LaTeX in R Markdown

I have added a link on Canvas in the resources section to help find LaTeX symbols

How else could you get help to learn more about LaTeX symbols?

To repeat: avoid hard to debug code!

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

If you document isn't knitting:

- **For code:** use the # symbol until you can find the line of code that is giving the error message
- **For syntax:** cut part of the document until it knits and then paste it back

Practice homework 0

To gain practice with R and R Markdown in preparation for homework 1, we have created a practice homework 0.

To download the homework please do the following:

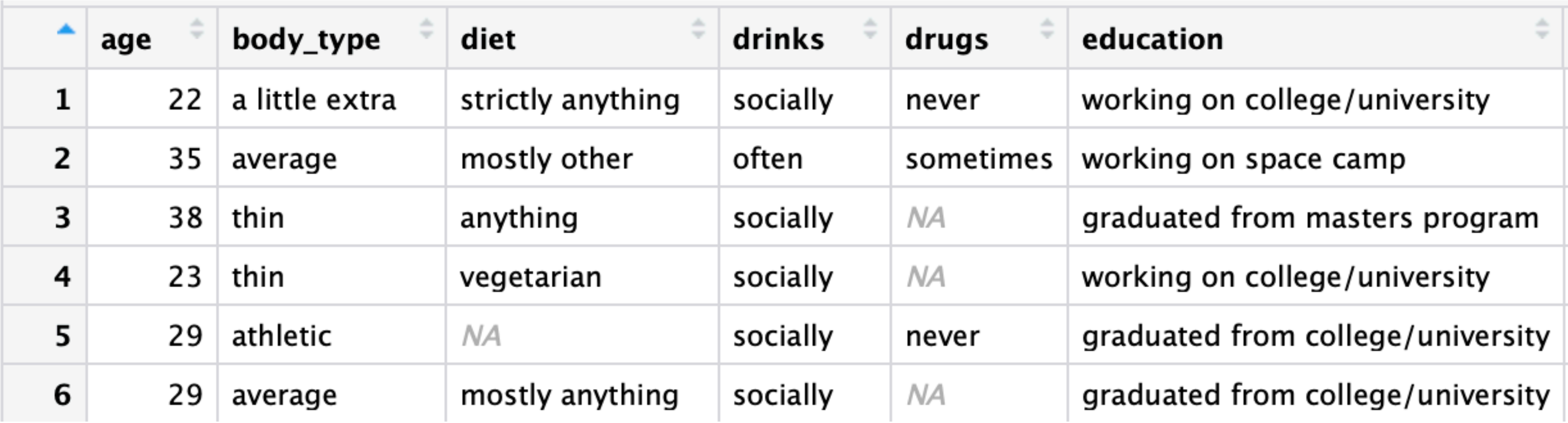
```
> library(SDS230)
> download_homework(0)
```

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

Questions?

Data frames

Data frames contain structured data

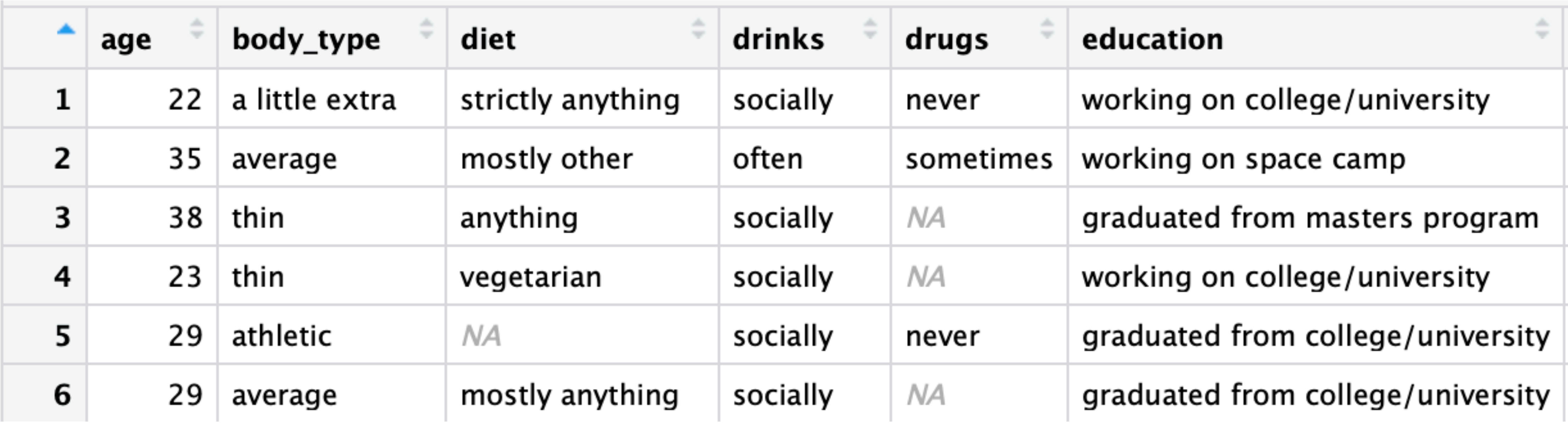


	age	body_type	diet	drinks	drugs	education
1	22	a little extra	strictly anything	socially	never	working on college/university
2	35	average	mostly other	often	sometimes	working on space camp
3	38	thin	anything	socially	NA	graduated from masters program
4	23	thin	vegetarian	socially	NA	working on college/university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

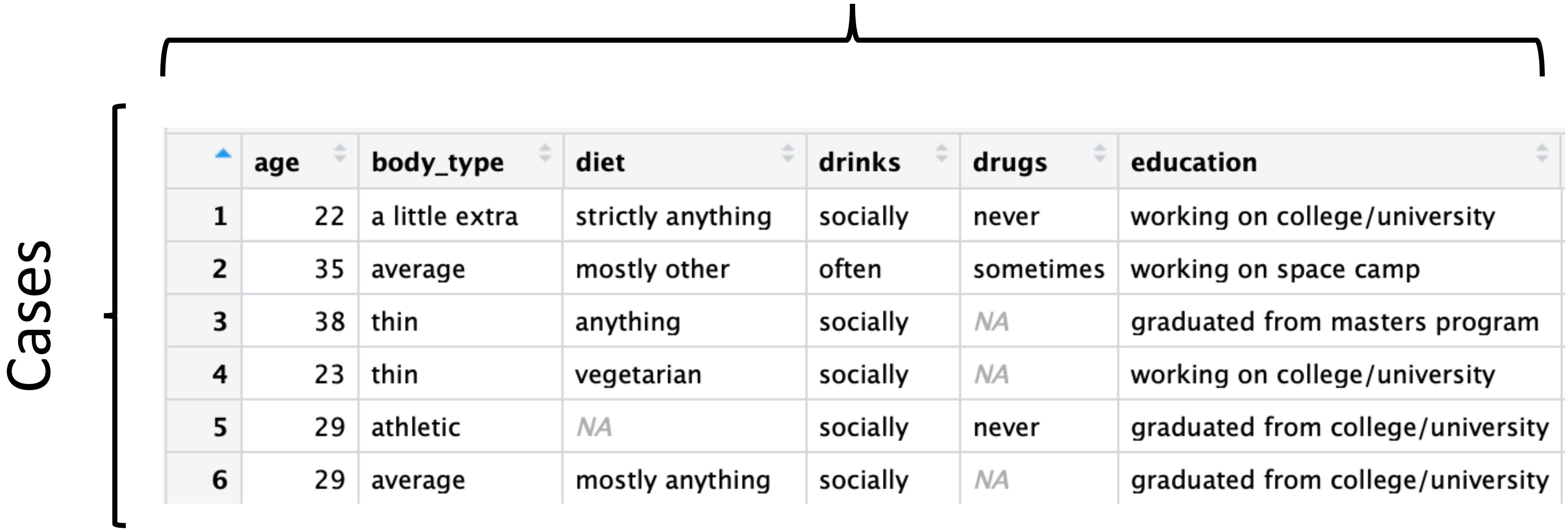
Data frames contain structured data

```
> install.packages("okcupiddata")      #  only needs to be run once
> library(okcupiddata)
> 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

Cases

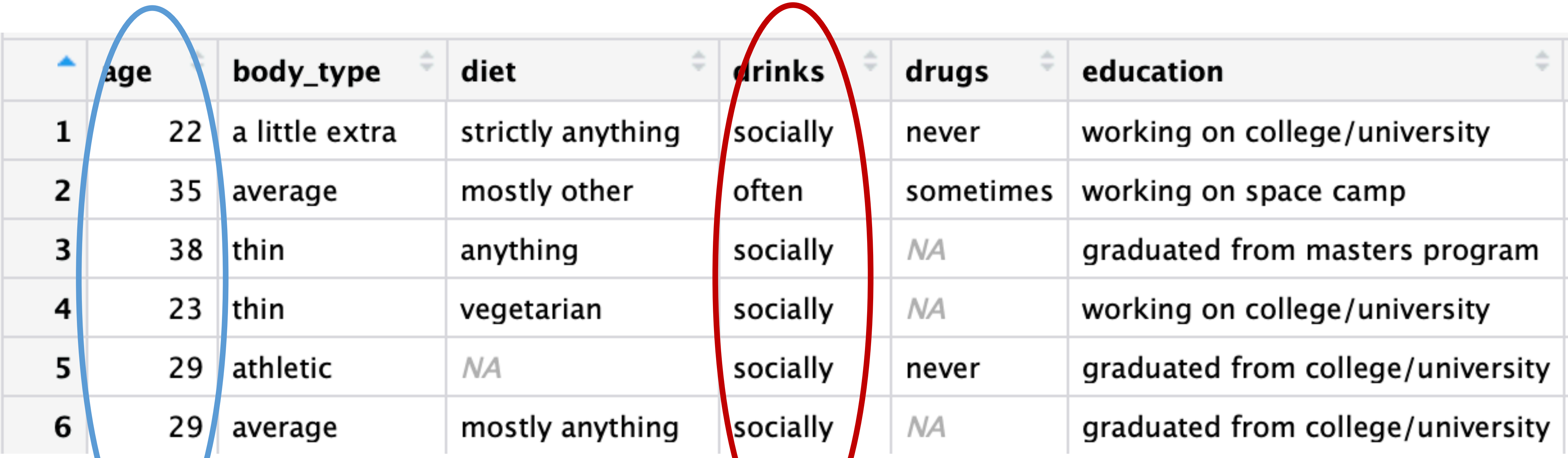
	age	body_type	diet	drinks	drugs	education
1	22	a little extra	strictly anything	socially	never	working on college/university
2	35	average	mostly other	often	sometimes	working on space camp
3	38	thin	anything	socially	NA	graduated from masters program
4	23	thin	vegetarian	socially	NA	working on college/university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

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

When data is loaded from a package it isn’t visible in the environment pane. We can make it visible using the data() function.

```
> library(okcupiddata)
> data(profiles)
```

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

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

Data frames

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

Can you get the mean() age of users in this data set?

Extracting rows from a data frame

We can extract rows from a data frame in a similar way as extracting values from a vector by using the square brackets

```
> profiles[1, ]  # returns the first row of the data frame
> profiles[, 1]  # returns the first column of the data
```

Note, the first column of the profiles data frame is the variable *age*, so we can also get the first column using:

```
> profiles$age  # this is the same as profiles[, 1]
```

Extracting rows from a data frame

We can also create vectors of numbers or booleans specifying which rows we want to extract from a data frame

```
# create a vector with the numbers 1, 10, 20
> my_vec <- c(1, 10, 20)

# use my_vec to get the 1st, 10th, and 20th row in profiles
> small_profiles <- profiles[my_vec, ]
> dim(small_profiles)  # number of rows and columns in the data frame
```

Extracting rows from a data frame

Finally, we can also extract rows by creating a Boolean vector that is of the same length as the number of rows in the data frame

TRUE values will be extracted from the data frame while FALSE values will not

```
# create a vector of booleans
> my_bools <- c(TRUE, FALSE, TRUE)

# use the Boolean vector to get the 1st and 3rd row
> small_profiles[my_bools, ]
```

Questions?

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 heights 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

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

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)
```

What is the proper statistical notation for these values: $\hat{p}$ or π ?

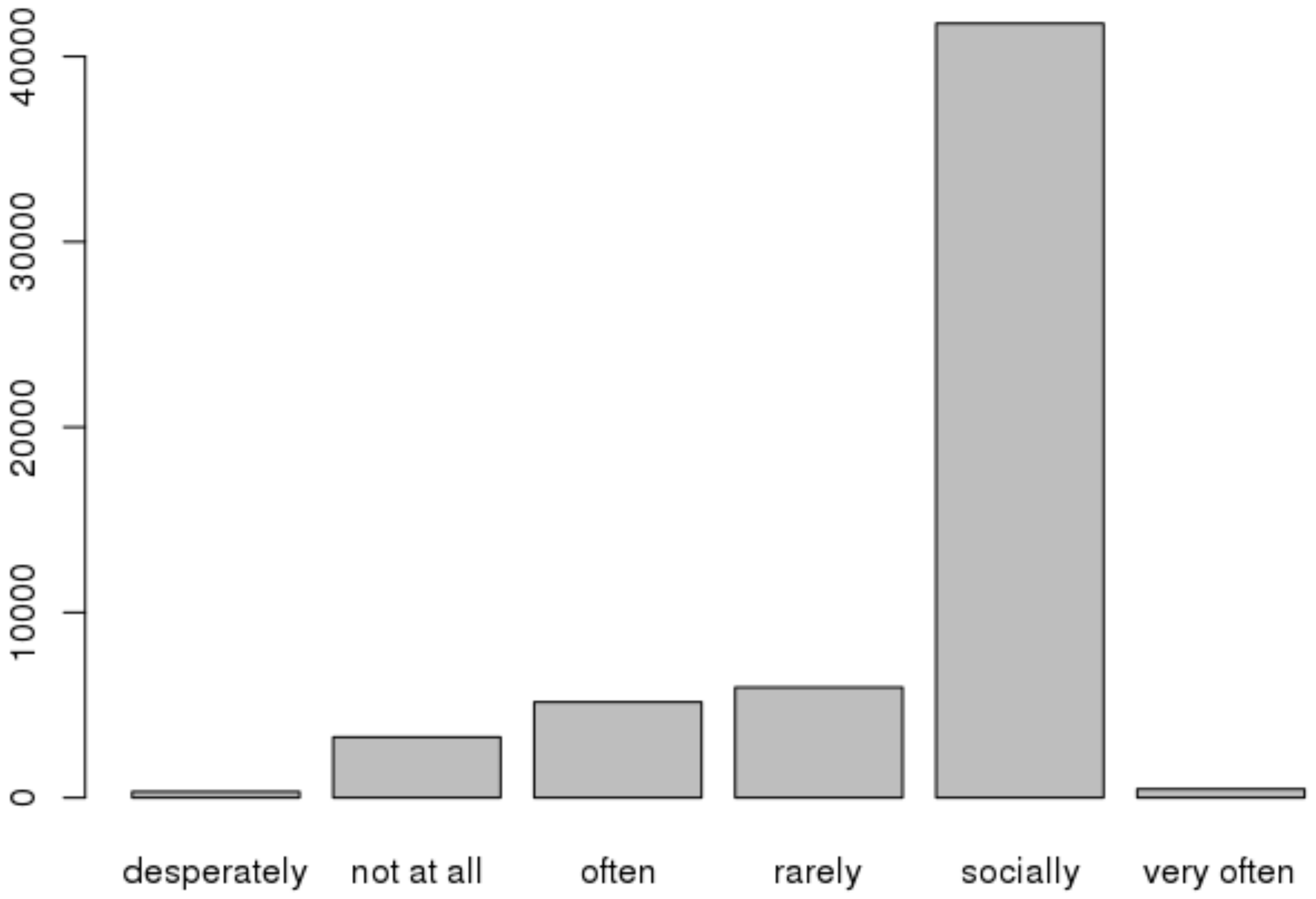
Bar plots

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)
```

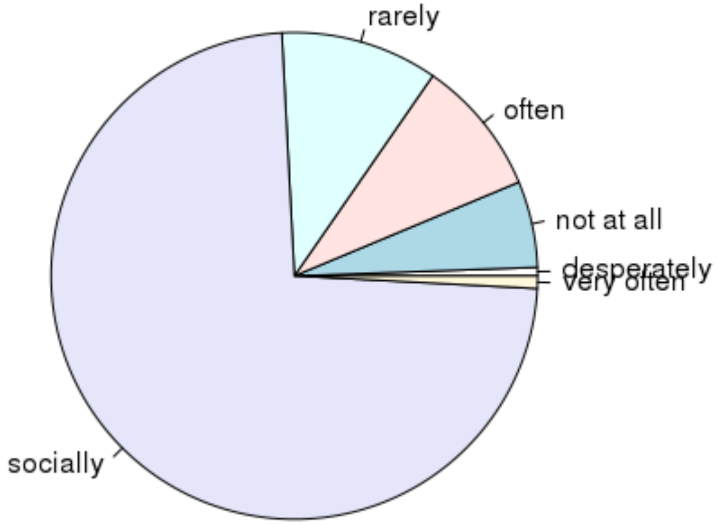


What is wrong with this plot?

Pie charts

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

```
> pie(drinks_table)
```



Which is best: bar plots or pie charts?

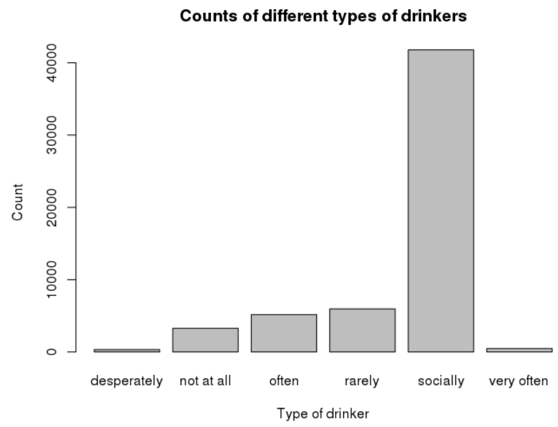
```
> barplot(table(profiles$sex, useNA = "always"))

> 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?

Removing social drinkers



Social drinkers are dominating our plot ☹

We can get rid of social drinkers by only plotting counts less than 10,000

```
> nonsocial_inds <- drinks_table < 10000
> nonsocial_drinks_table <- drinks_table[nonsocial_inds]
> barplot(nonsocial_drinks_table)
```

Questions?

Quantitative data: statistics

There are several statistics that describe the central tendency of quantitative data?

- The mean: mean()
- The median: median()

Which of these measures is robust to outliers?

Can you calculate the mean and median of OkCupid user's heights?

What went wrong?

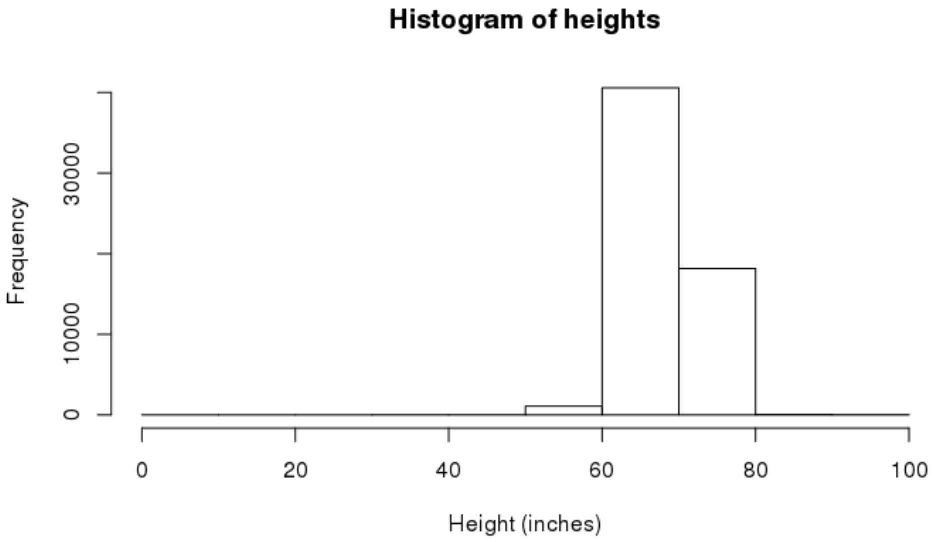
What is the proper statistical notation for the mean of OkCupid user's heights:
$\bar{x}$ or μ ?

Quantitative data: Visualizing heights

Q: How can we visualize the heights in the profiles data frame?

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

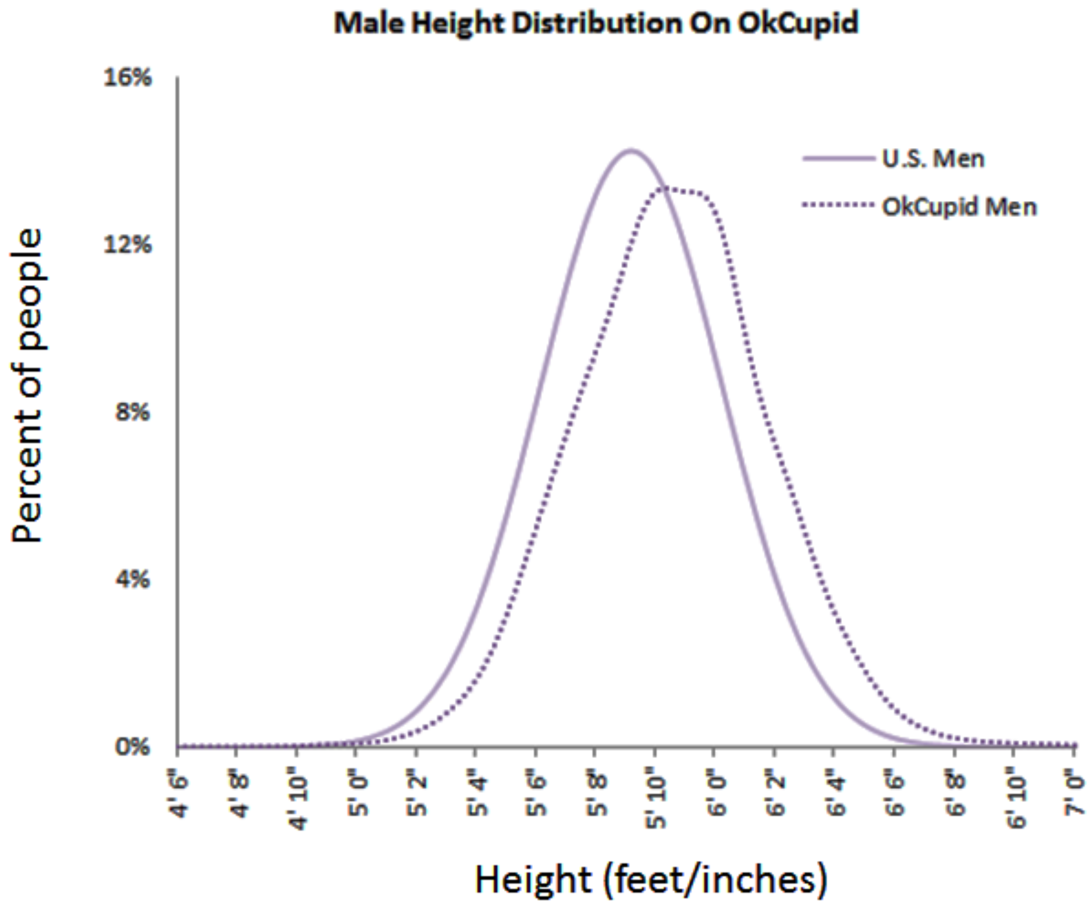
We can create histograms in R using the hist() function

Can you create a histogram of heights?

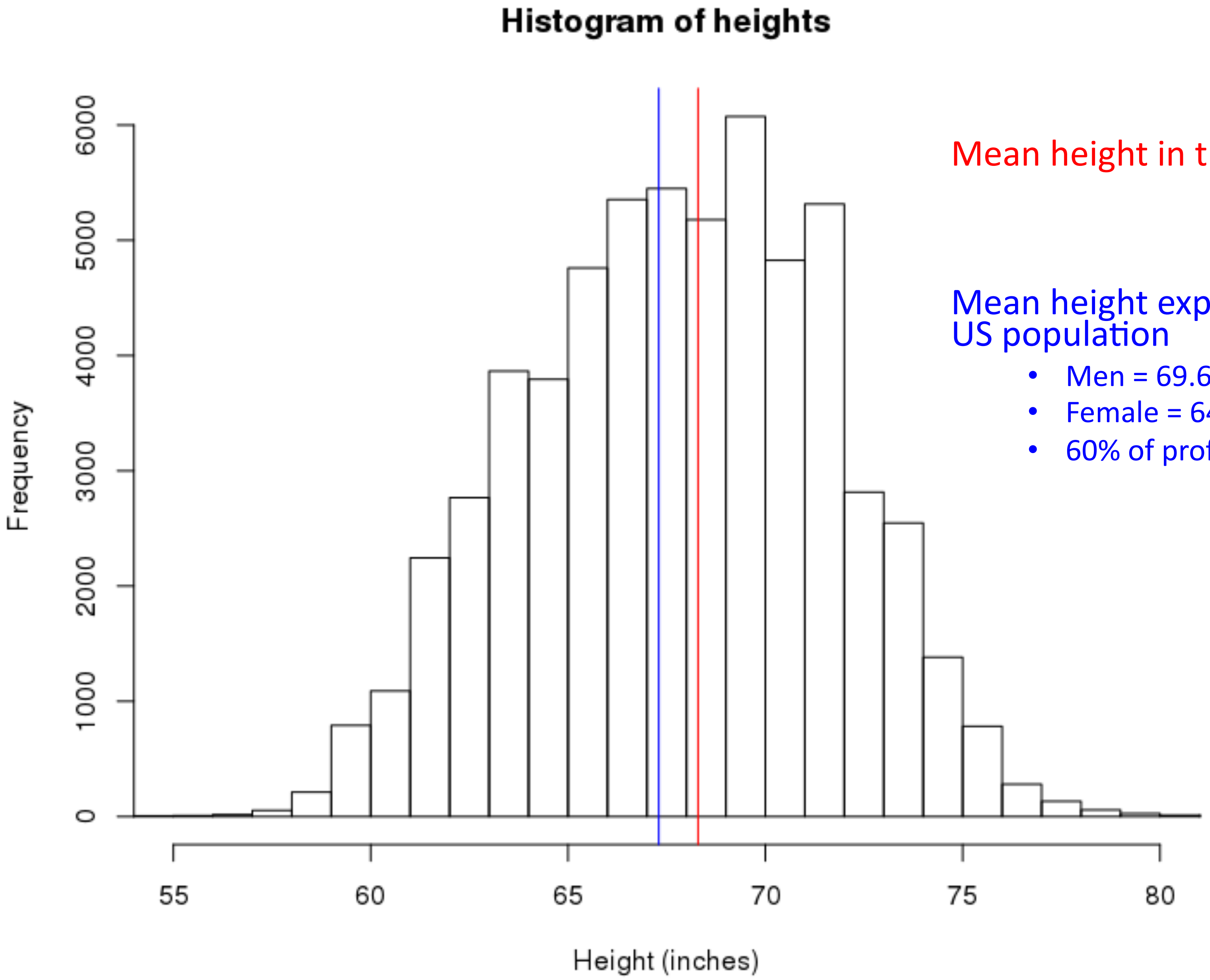
```
> hist(profiles$height)

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

OkCupid users are taller than the average person



Can we see this in the profiles data?



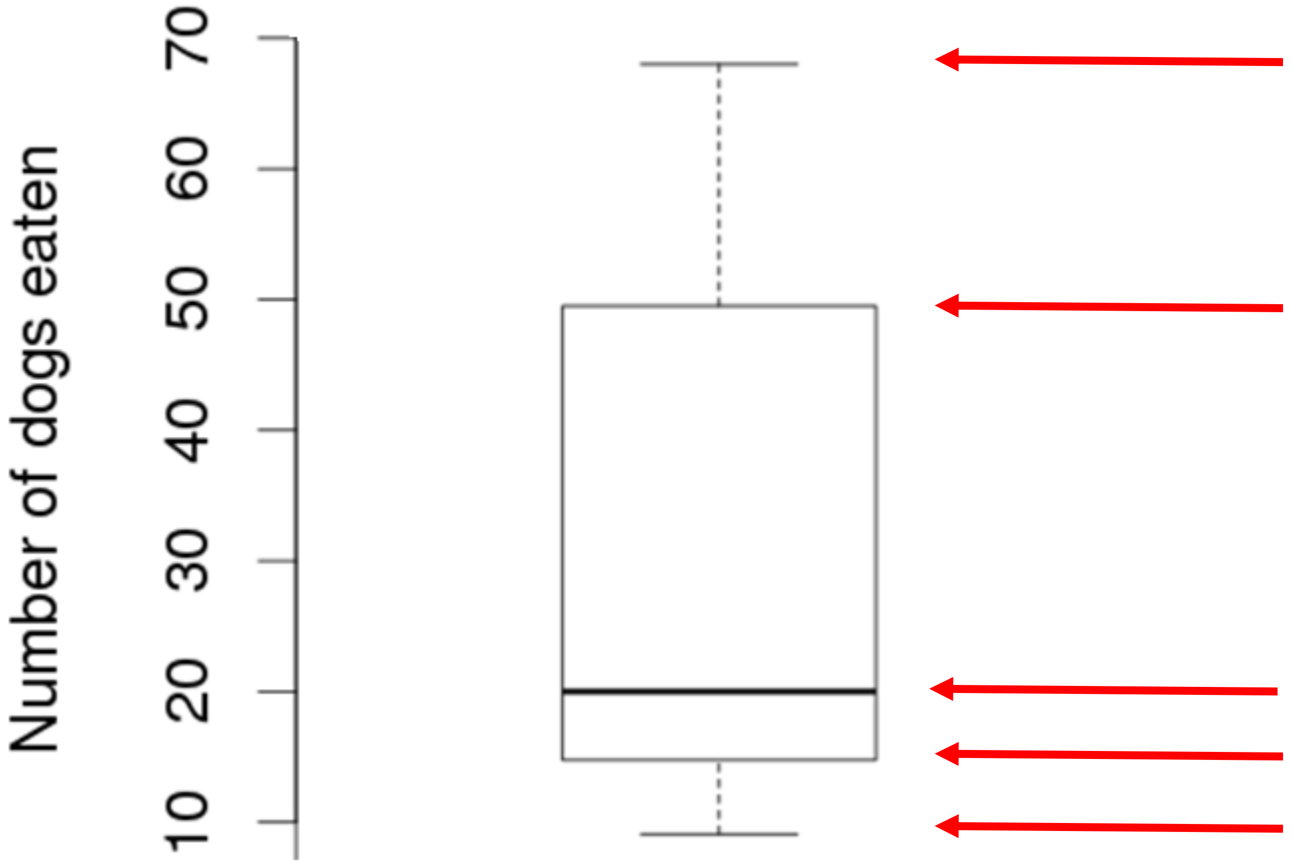
Mean height in the profiles data

Mean height expected based on US population

- Men = 69.6”
- Female = 64”
- 60% of profiles are men

abline() adds lines to plots

Box plots can also visualize quantitative data



Maximum value that is not an outlier

Q_3

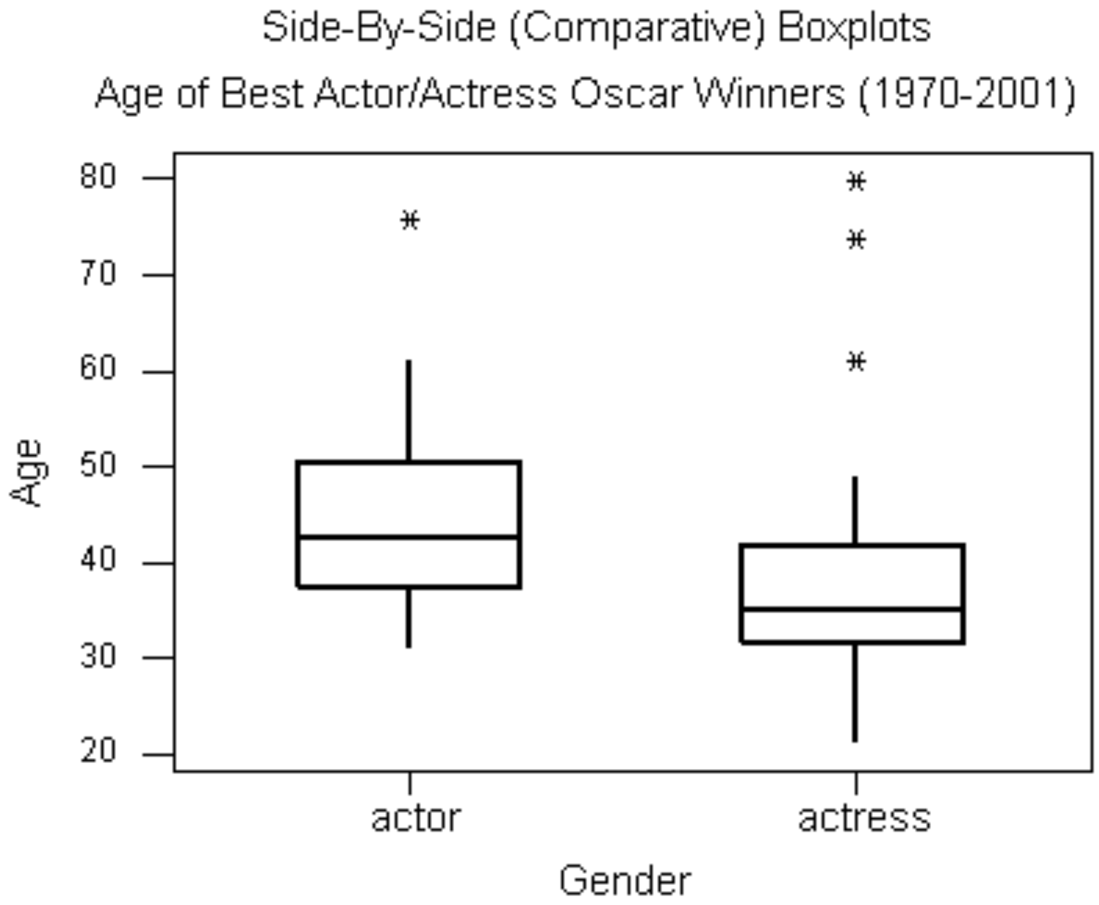
Median

Q_1

Minimum value that is not an outlier

```
R: boxplot(v)
```

Side-by-side boxplots



Useful for comparing distributions!

- What does the figure above show?

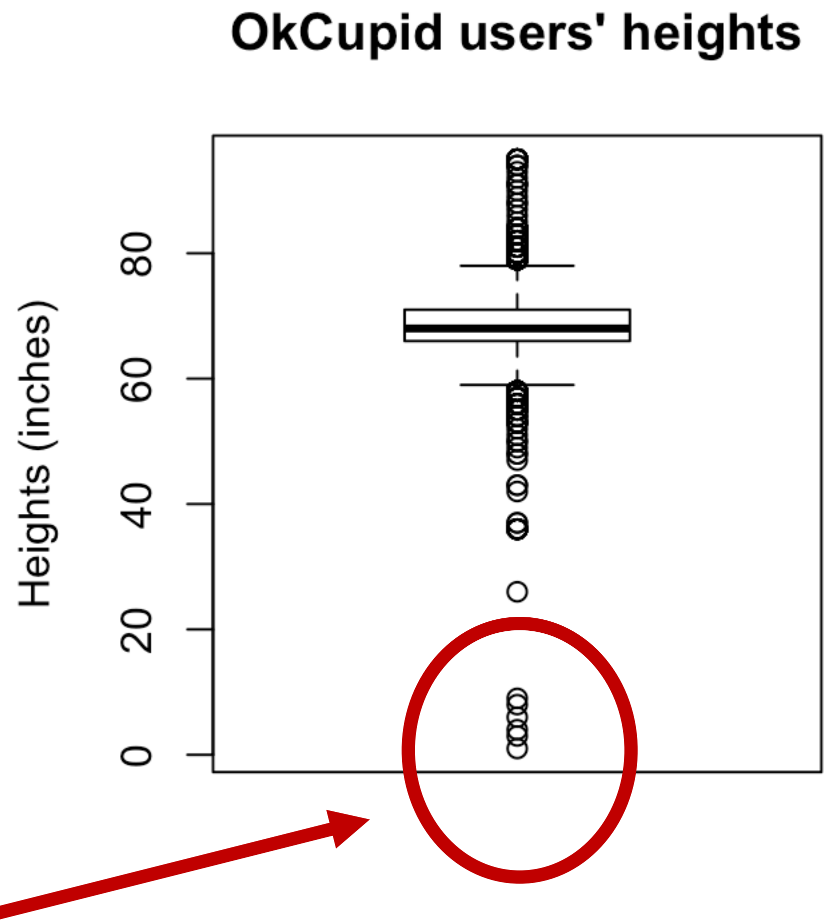
Outliers

Outliers on boxplots are values that are more than 1.5 * IQR

What should we do if we have outliers?

Investigate!

- If there are due to an error, remove them



People under 20” tall?

Outliers

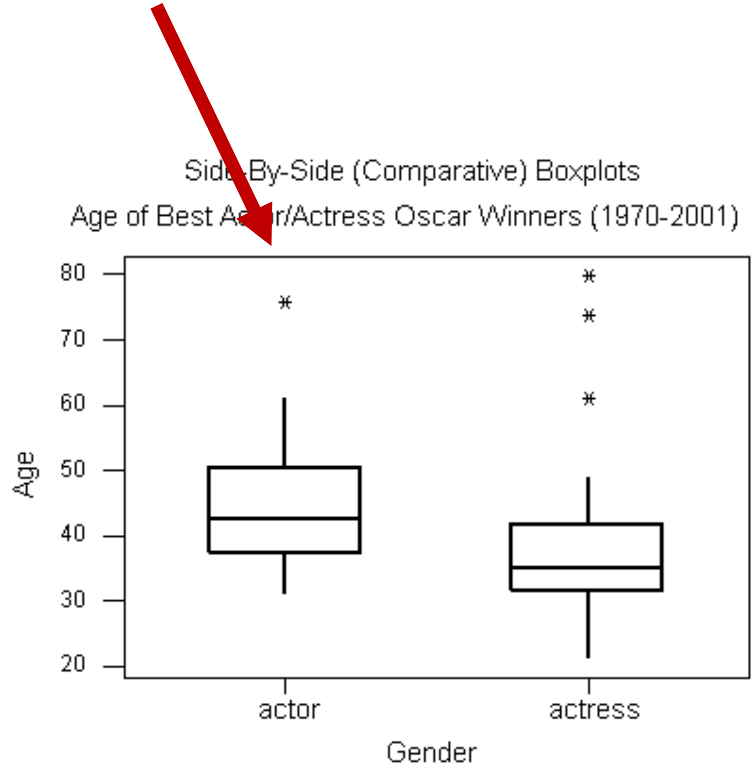
Outliers on boxplots are values that are more than 1.5 * IQR

What should we do if we have outliers?

Investigate:

- If there are due to an error, remove them
- **If not, need to account for them**

Who is this actor?



Questions?

CitiBike data

Let’s look at the bike share data from NYC

```
> load('daily_bike_totals.rda’)
```

CitiBike analysis

What does each case correspond to?

We can use the dim() function to get how many cases and variables there are

- How many are there?

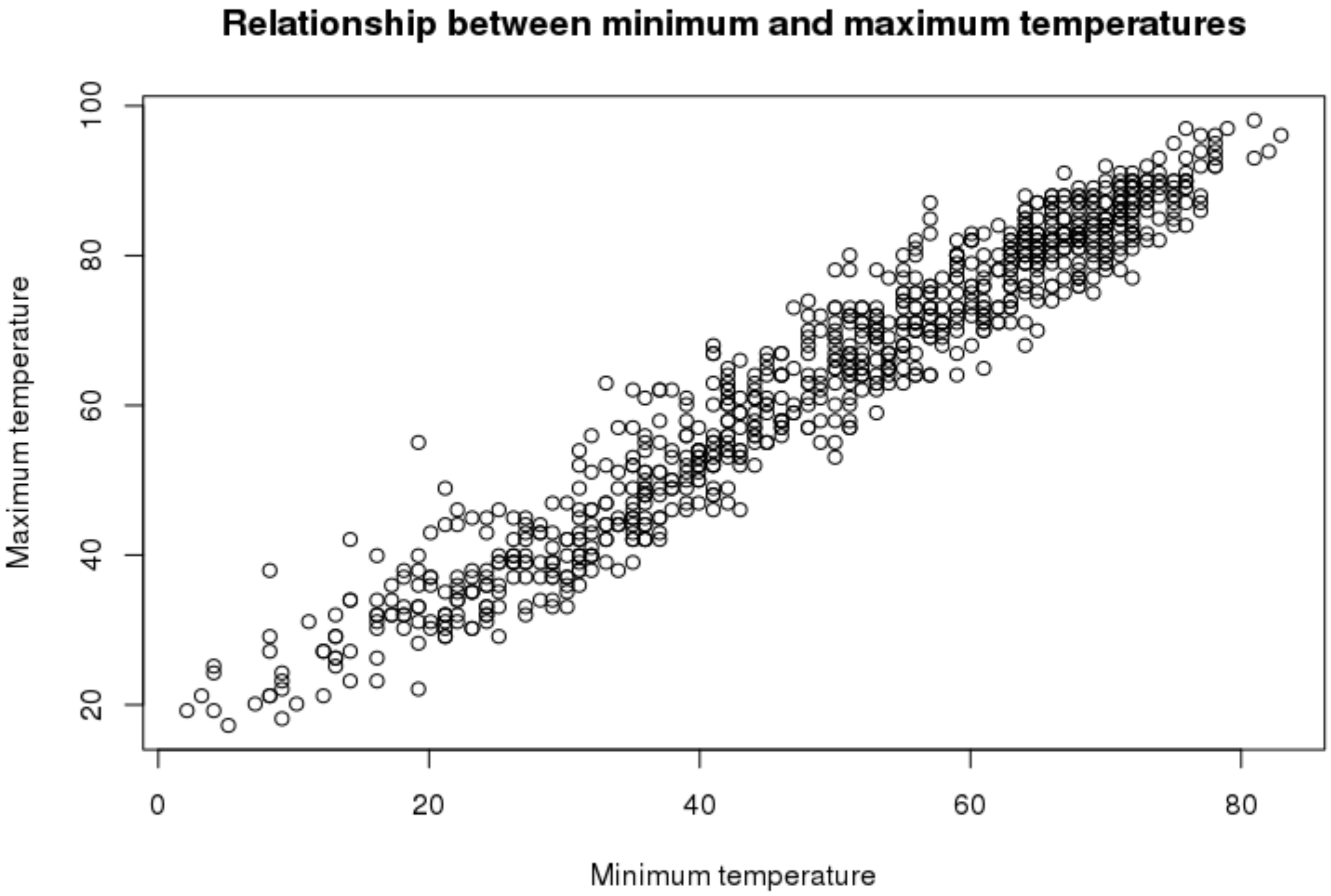
Scatter plots

We can use the plot(x, y) function to create scatter plots

Can you create a scatter plot of the relationship between the minimum and maximum temperatures?

```
> plot(bike_daily_data$min_temperature,
        bike_daily_data$max_temperature,
        xlab = "Minimum temperature",
        ylab = "Maximum temperature",
        main = "Relationship between min and temp")
```

Scatter plots



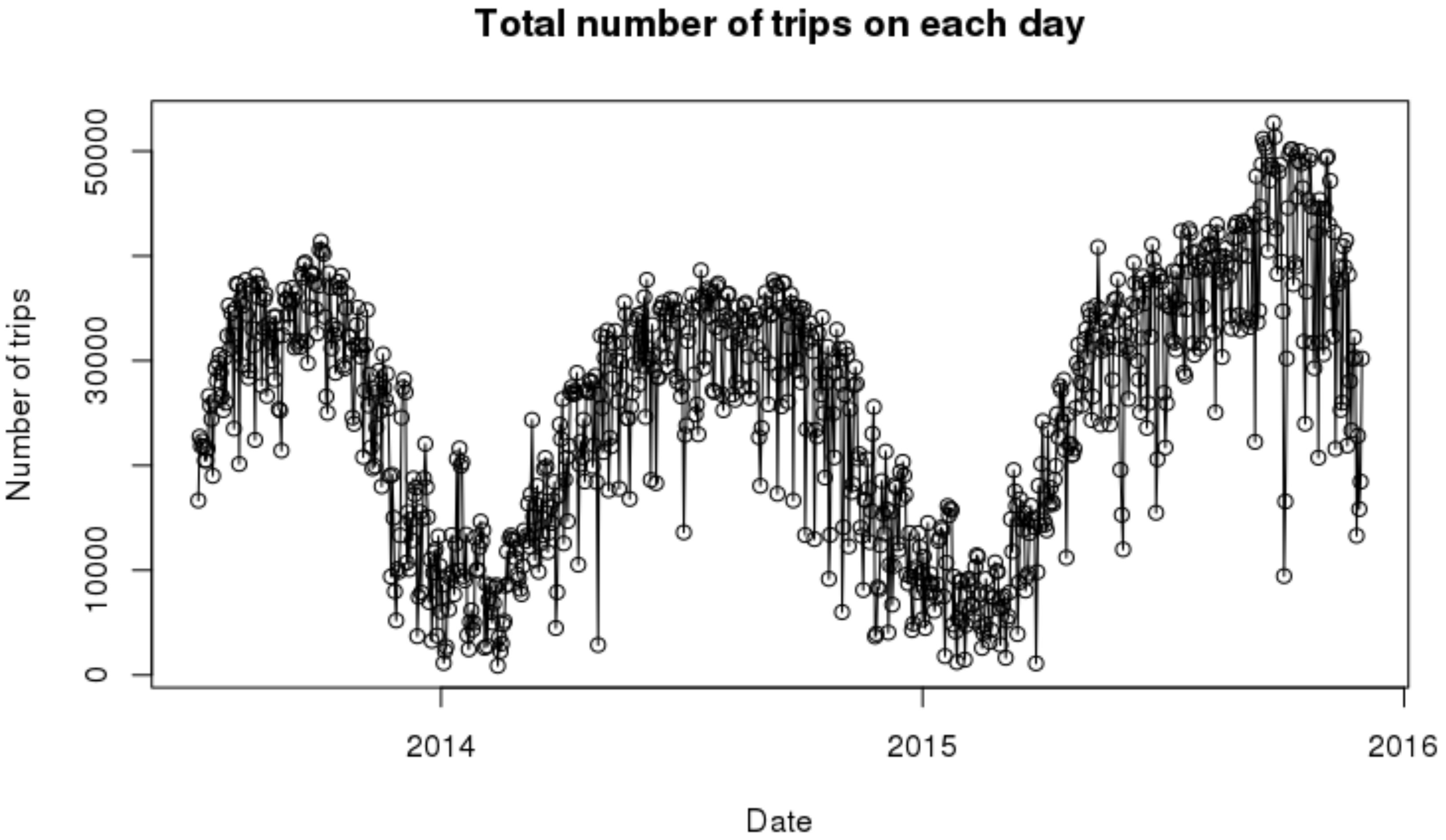
Plotting time series

We can use the plot(x, y) function to plot time series

```
# we can connect the points in a plot using
> plot(x, y, type = 'l')    # connected points
> plot(x, y, type = 'o')   # both points and dots

> plot(bike_daily_data$date,  bike_daily_data$trips,
            type = 'o',
            xlab = "Date",
            ylab = "Number of trips",
            main = "Total number of trips on each day")
```

Plotting time series



For next class

Try the practice homework:

```
> library(SDS230)
> download_homework(0)
```

If you want to practice R more, try the intro to R DataCamp course

- I will post a link to this on Canvas