Plots and statistics for categorical and quantitative data

Overview

Review of vectors and continuation of data frames

Statistics and plots for categorical data in R

Statistics and plots for quantitative data in R

Announcement: learning groups!

Stephan is organizing learning groups where students can get together (independent of TAs) to work on the homework and other class projects.

If you are interested in being part of a learning group, please sign up by midnight tonight (Thursday).

- A link to sign up is on Canvas and was sent out as an announcement.

There will also be two TA sessions to review R covered in class

- ~~Nathan: Wednesday evening at 7pm in Bass L01-A and on Zoom~~
- Amanda: Thursday at 4pm in Rosenkranz 302 and on Zoom

Announcement: Homework 1

Homework 1: SDS230::download_homework(1)

Due on Gradescope by 11pm on Sunday September 11th

- 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!

Review: vectors

Creating vectors

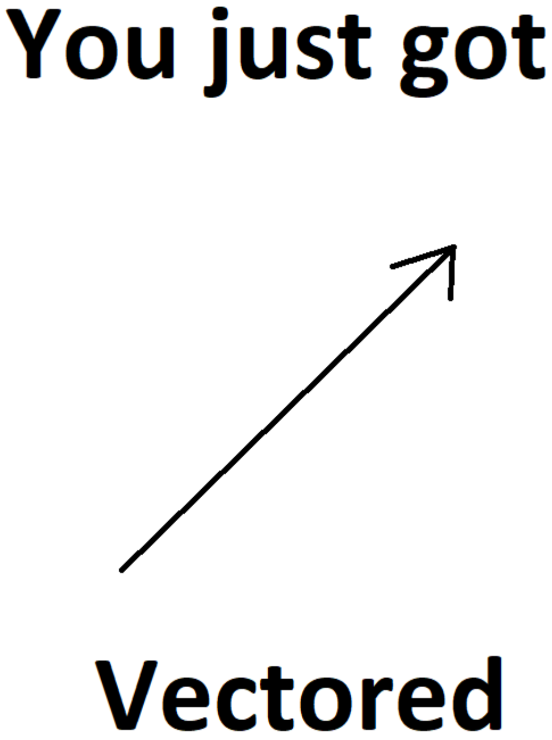
```
> s  <-  c("statistics", "data", "science", "fun")
> z <- 2:10
```

Accessing elements of vectors

```
> s[4]
> s[c(1, 2)]
```

Applying functions to vectors

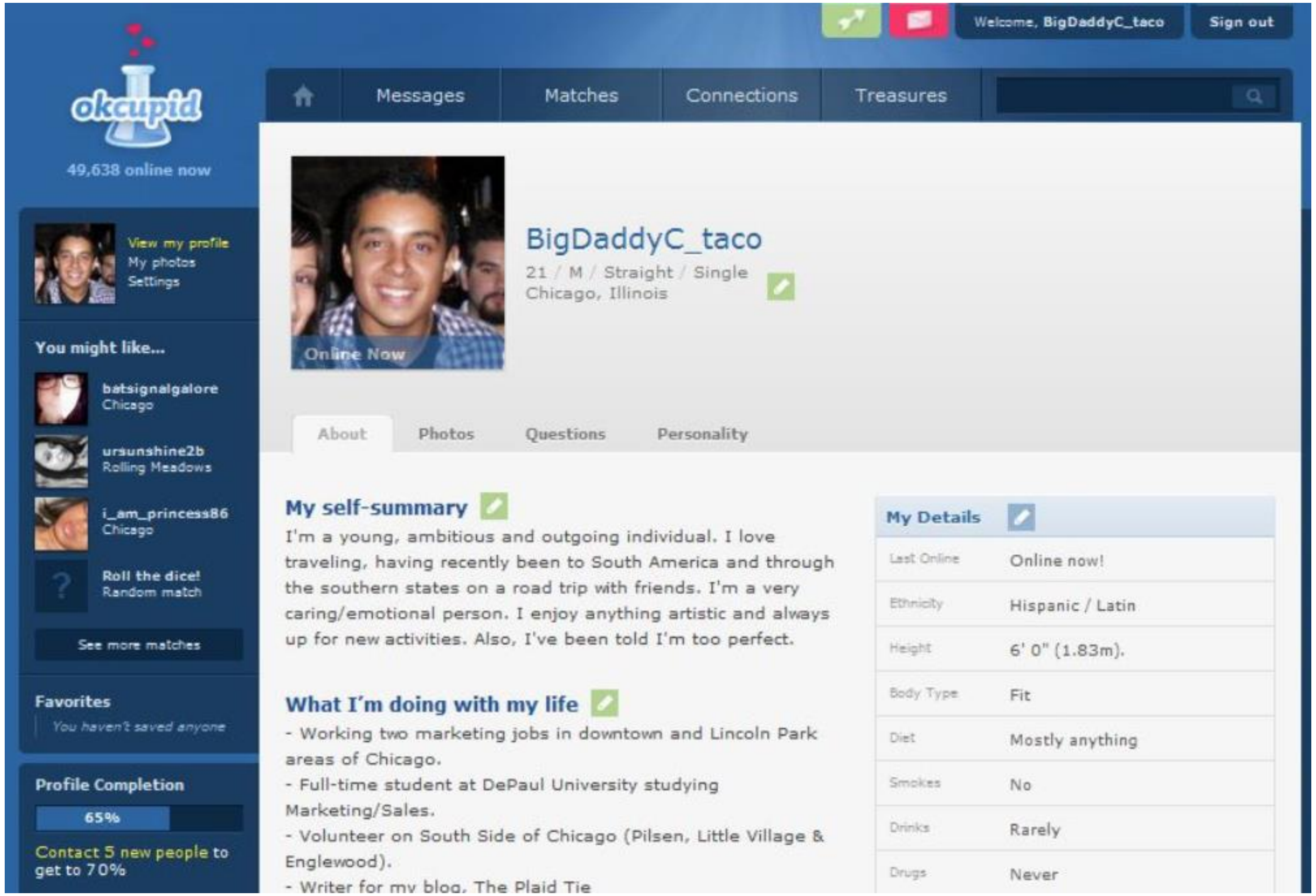
```
> sqrt(z)
> mean(z)
> z > 3
```



You just got

Vectored

OKCupid data



BigDaddyC_taco

21 / M / Straight / Single
Chicago, Illinois

My self-summary

I'm a young, ambitious and outgoing individual. I love traveling, having recently been to South America and through the southern states on a road trip with friends. I'm a very caring/emotional person. I enjoy anything artistic and always up for new activities. Also, I've been told I'm too perfect.

What I'm doing with my life

- Working two marketing jobs in downtown and Lincoln Park areas of Chicago.
- Full-time student at DePaul University studying Marketing/Sales.
- Volunteer on South Side of Chicago (Pilsen, Little Village & Englewood).
- Writer for my blog, The Plaid Tie

My Details	
Last Online	Online now!
Ethnicity	Hispanic / Latin
Height	6' 0" (1.83m).
Body Type	Fit
Diet	Mostly anything
Smokes	No
Drinks	Rarely
Drugs	Never

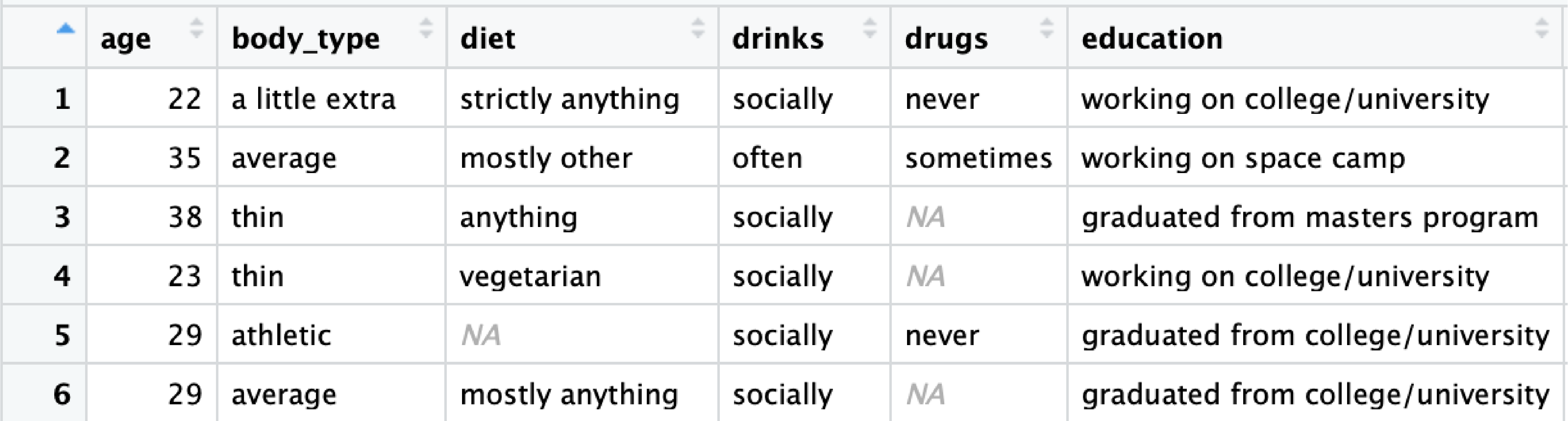
Did everyone read the article The Big Lies People Tell in Online Dating?

We are a little behind schedule so we will skip a discussion on this, but if you haven't done so already, please read article and fill out the survey about it

Data frames

Data frames contain structured data

```
> library(SDS230)
> 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!
```

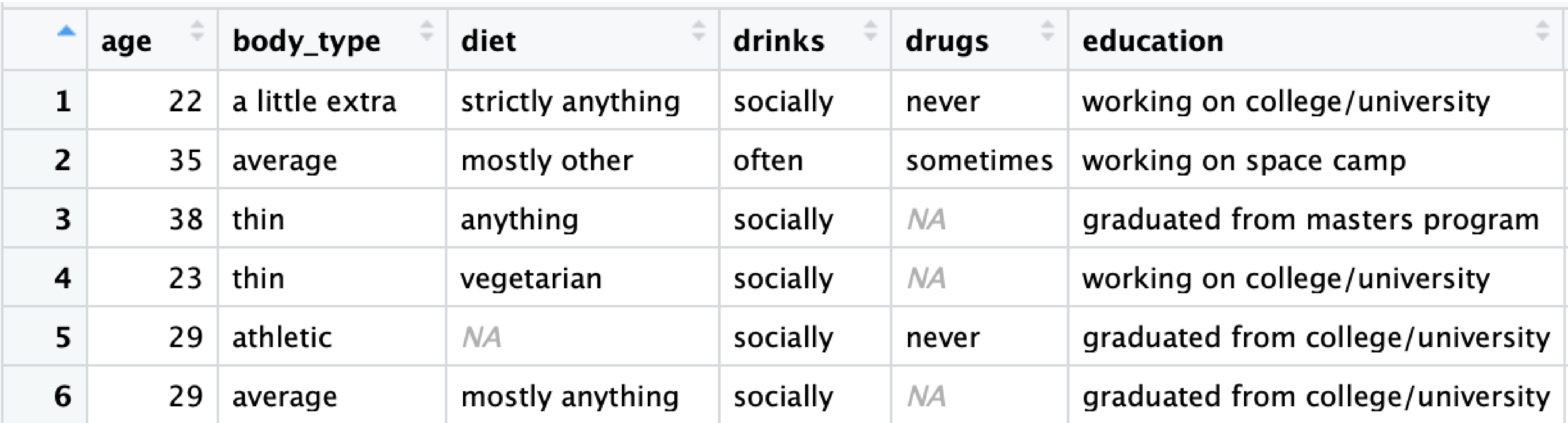


	age	body_type	diet	drinks	drugs	education
1	22	a little extra	strictly anything	socially	never	working on college/university
2	35	average	mostly other	often	sometimes	working on space camp
3	38	thin	anything	socially	NA	graduated from masters program
4	23	thin	vegetarian	socially	NA	working on college/university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

Data Frames

Variables

Cases



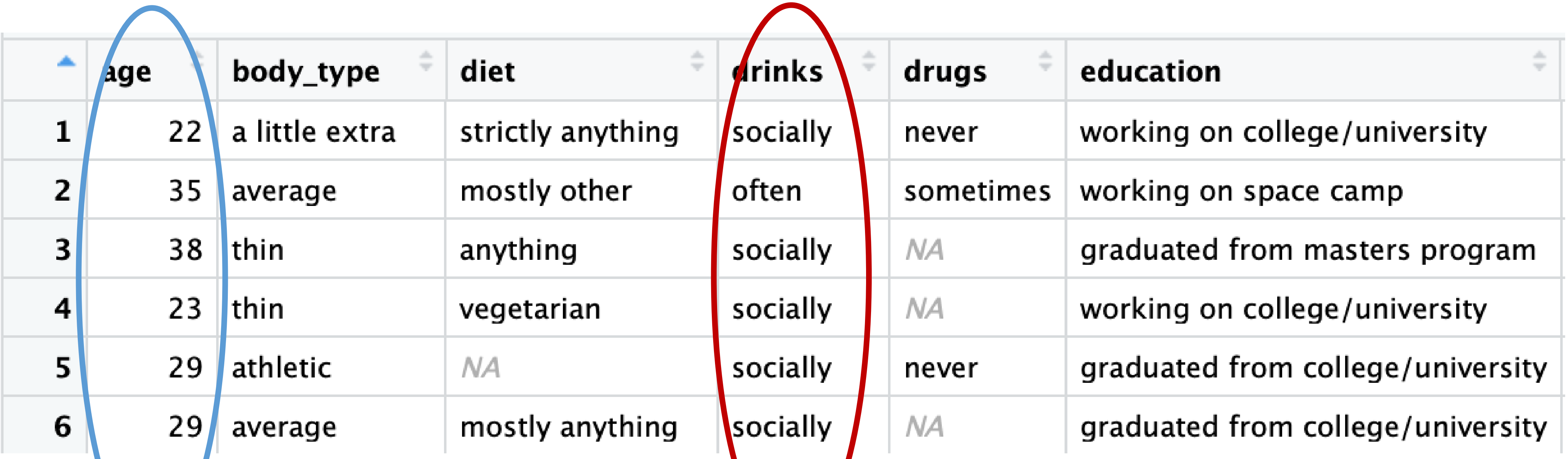
	age	body_type	diet	drinks	drugs	education
1	22	a little extra	strictly anything	socially	never	working on college/university
2	35	average	mostly other	often	sometimes	working on space camp
3	38	thin	anything	socially	NA	graduated from masters program
4	23	thin	vegetarian	socially	NA	working on college/university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

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

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

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

```
> mean(the_ages)
```

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 create vectors of numbers 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

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

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 ?

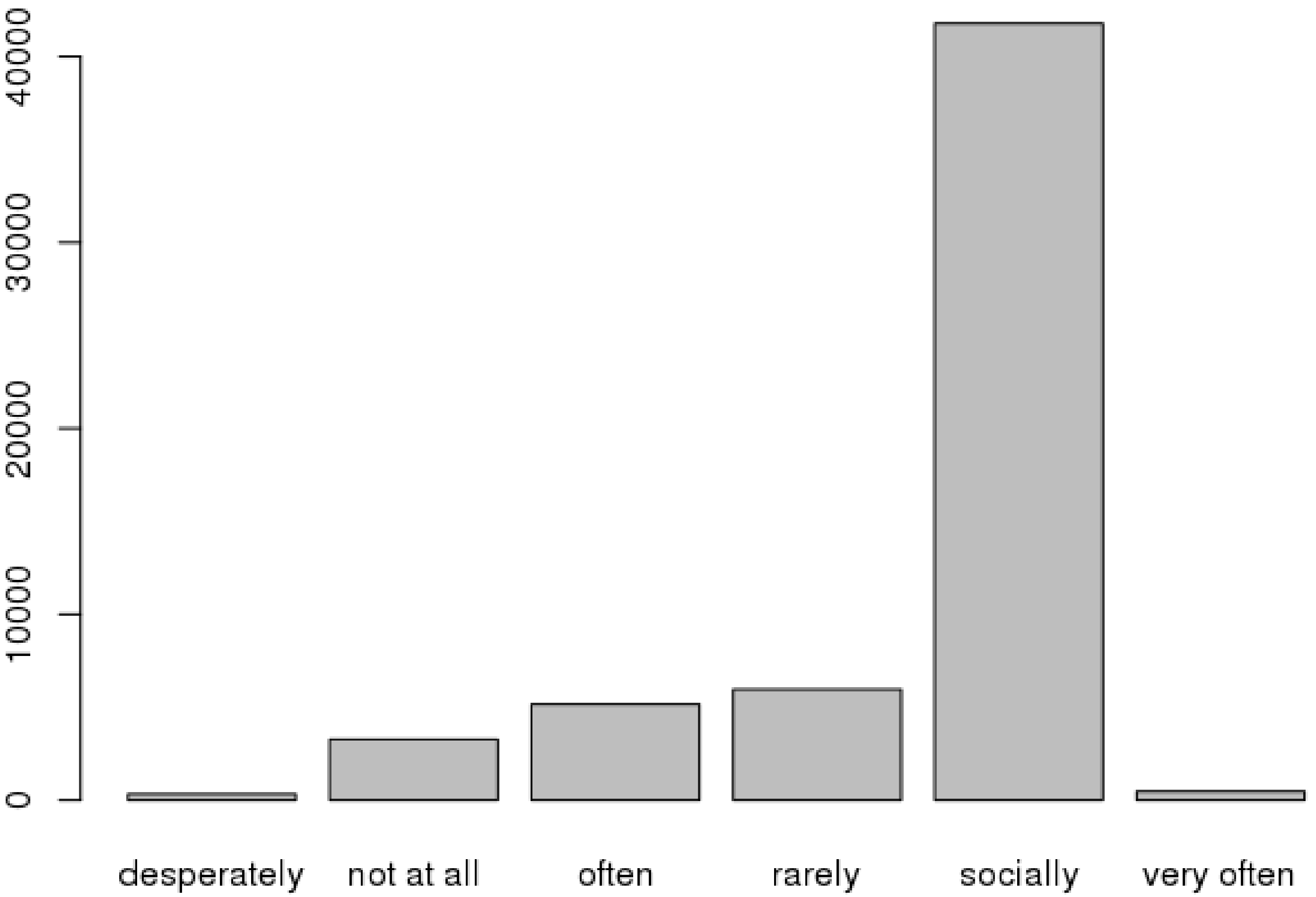
Bar plots (pun intended?)

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)
```



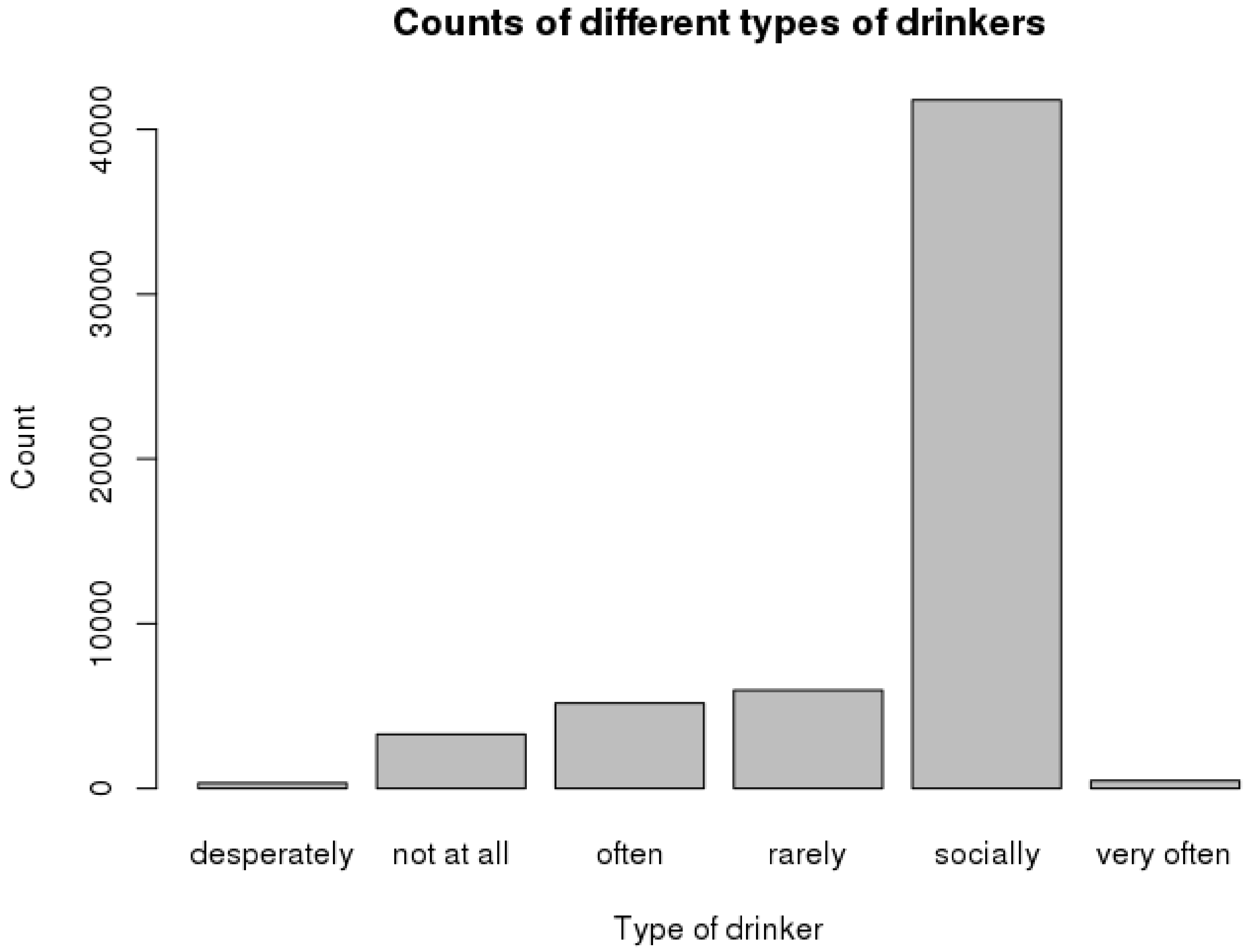
What is wrong with this plot?

Details matter!

Can you figure out how to label the axes?

- A: ? barplot
- A: xlab and ylab!

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

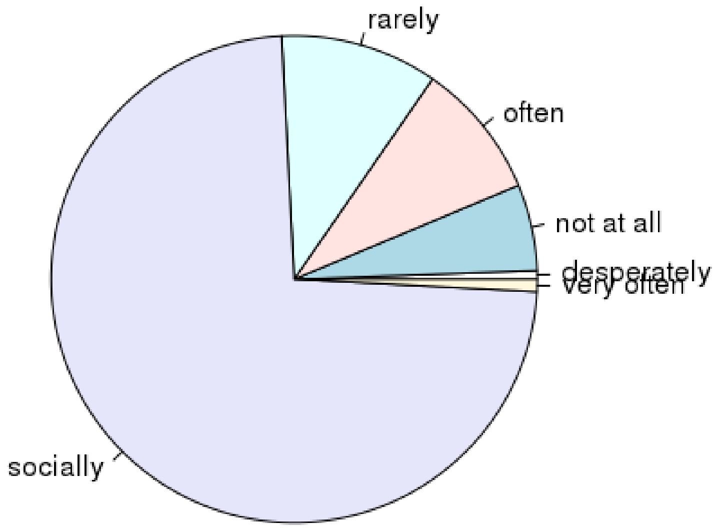


So much better!!!

Pie charts

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

```
> pie(drinks_table)
```



Which is best: bar plots or pie charts?

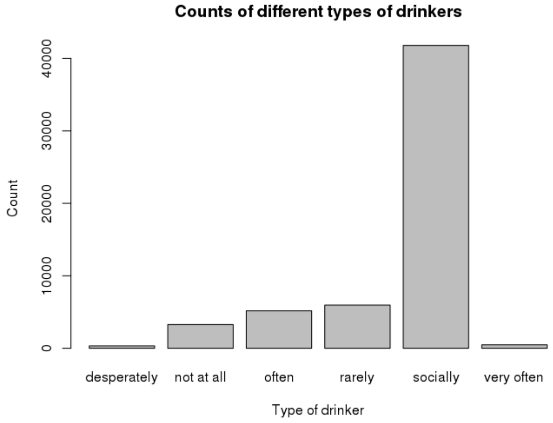
```
> barplot(table(profiles$sex, useNA = "always"))

> 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

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?

```
mean(v, na.rm = TRUE)
```

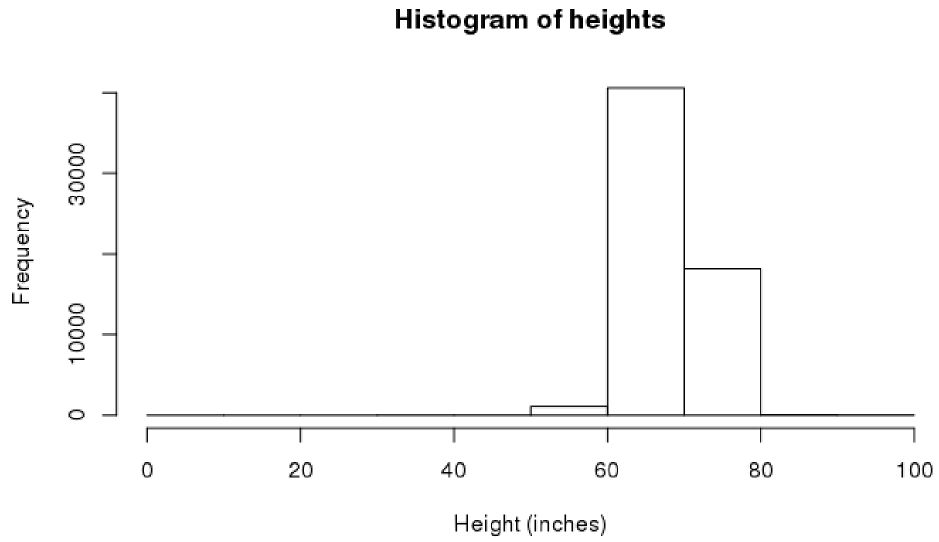
What is the proper statistical notation for the mean of OkCupid user's heights:
$\bar{x}$ or μ ?

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

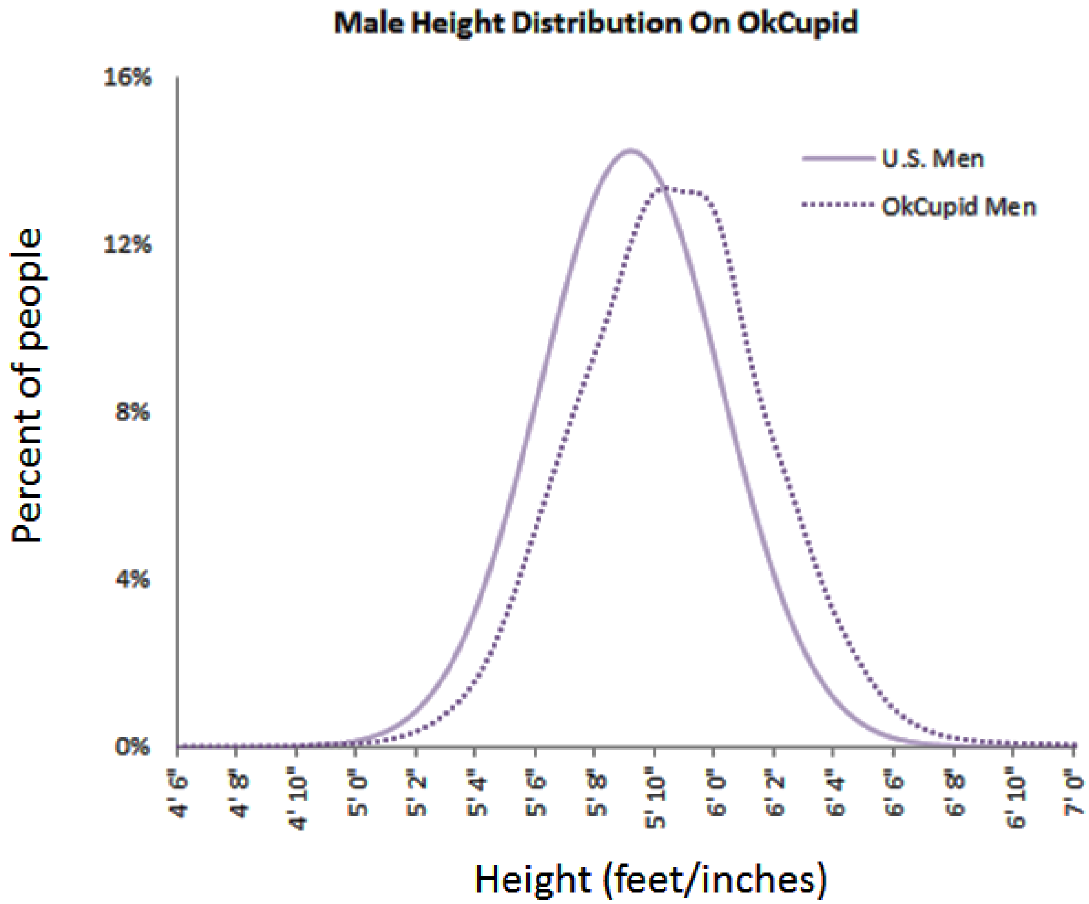
We can create histograms in R using the hist() function

Can you create a histogram of heights?

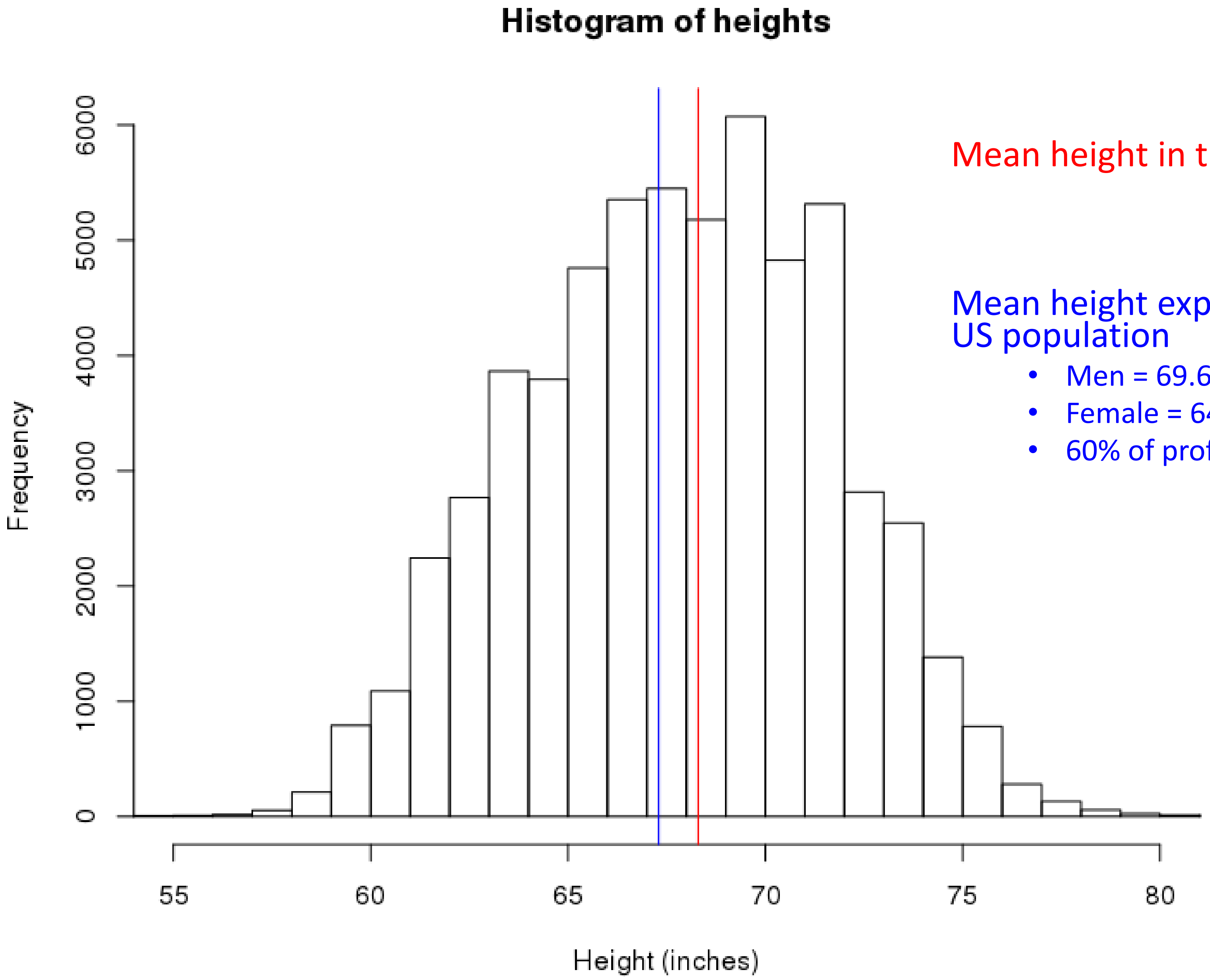
```
> hist(profiles$height)

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

OkCupid users are taller than the average person



Can we see this in the profiles data?



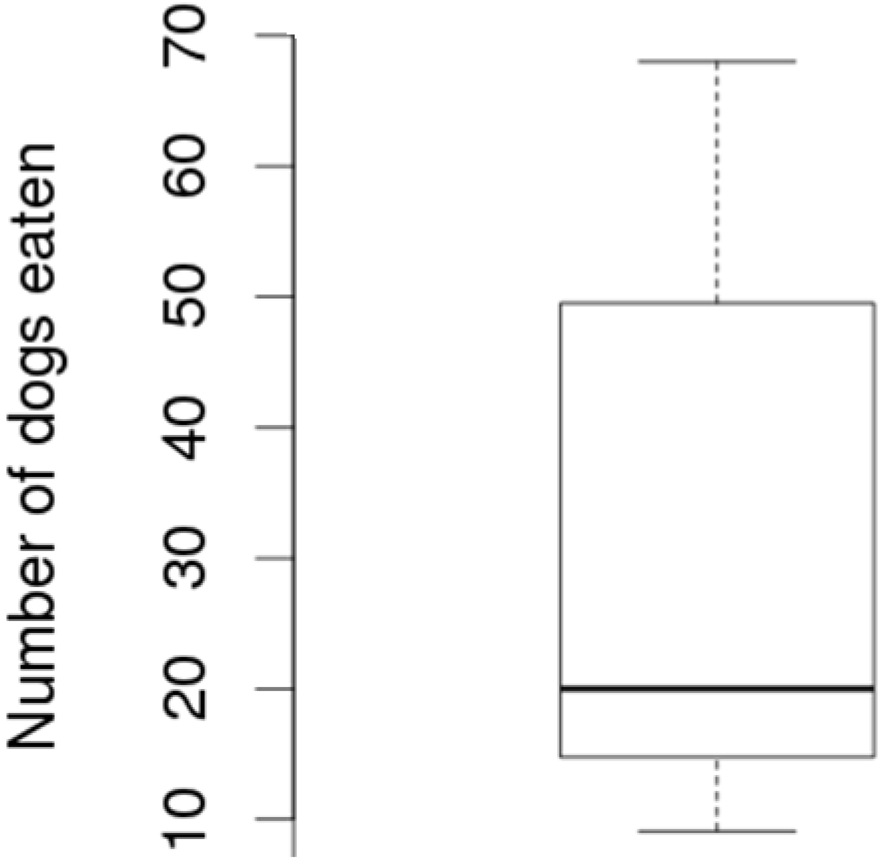
Mean height in the profiles data

Mean height expected based on US population

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

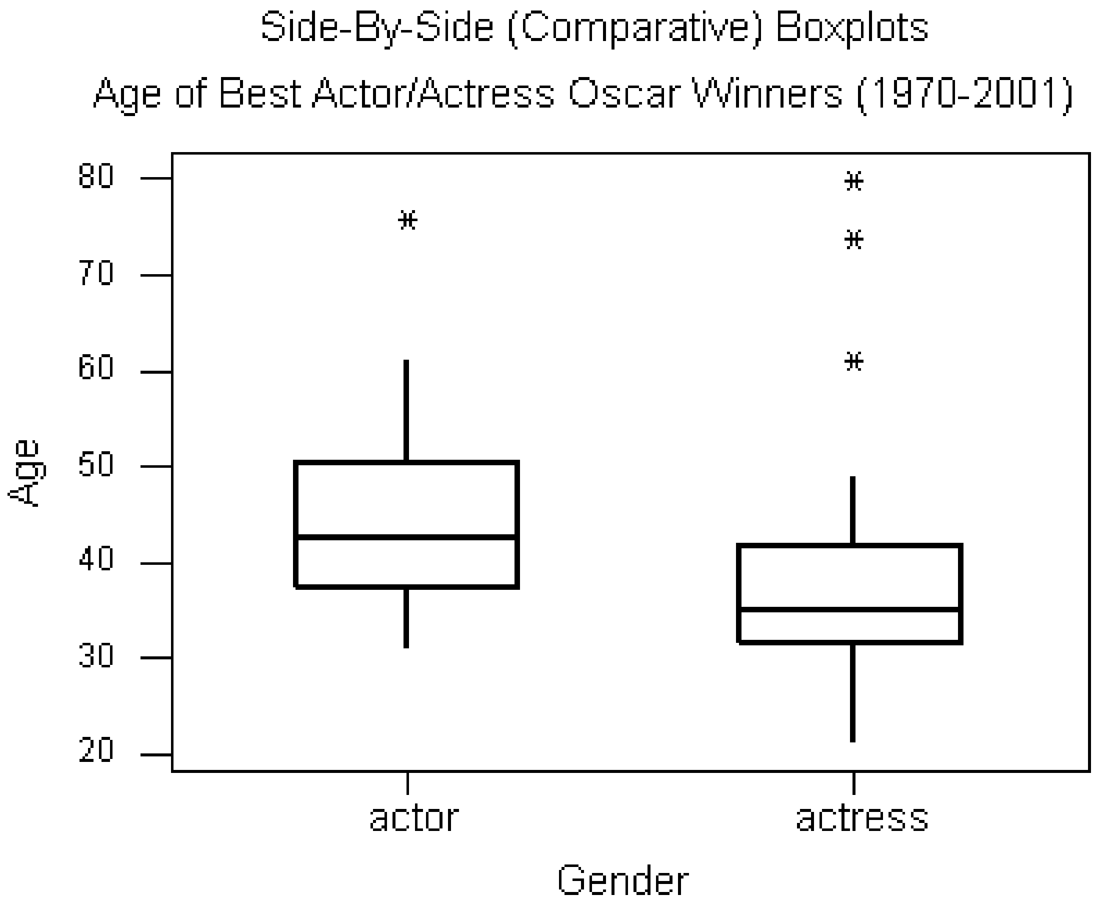
abline() adds lines to plots

Box plots can also visualize quantitative data



R: boxplot(v)

Side-by-side boxplots



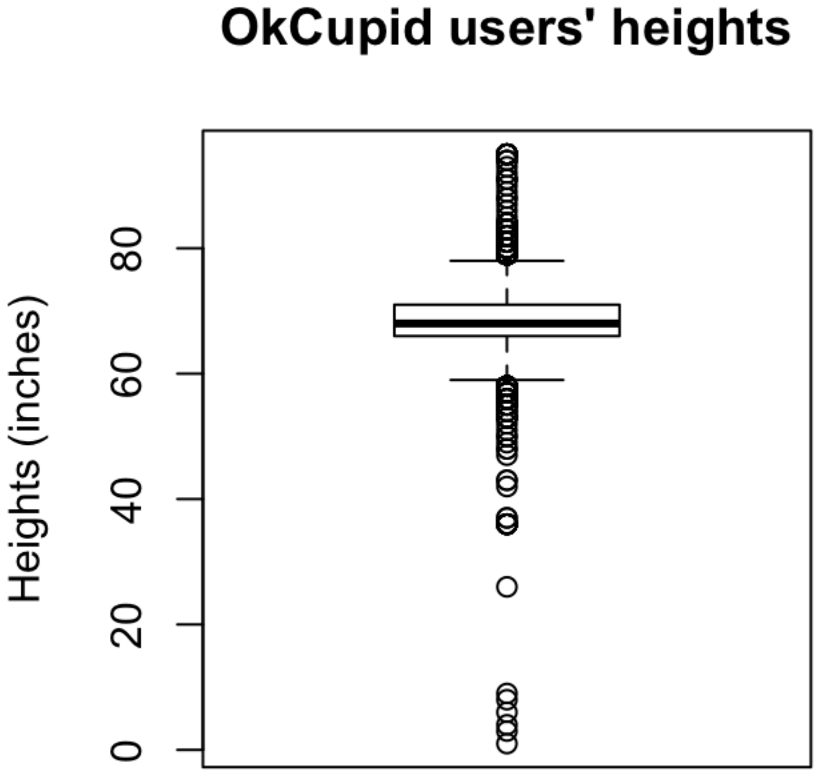
Useful for comparing distributions!

- 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?



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?

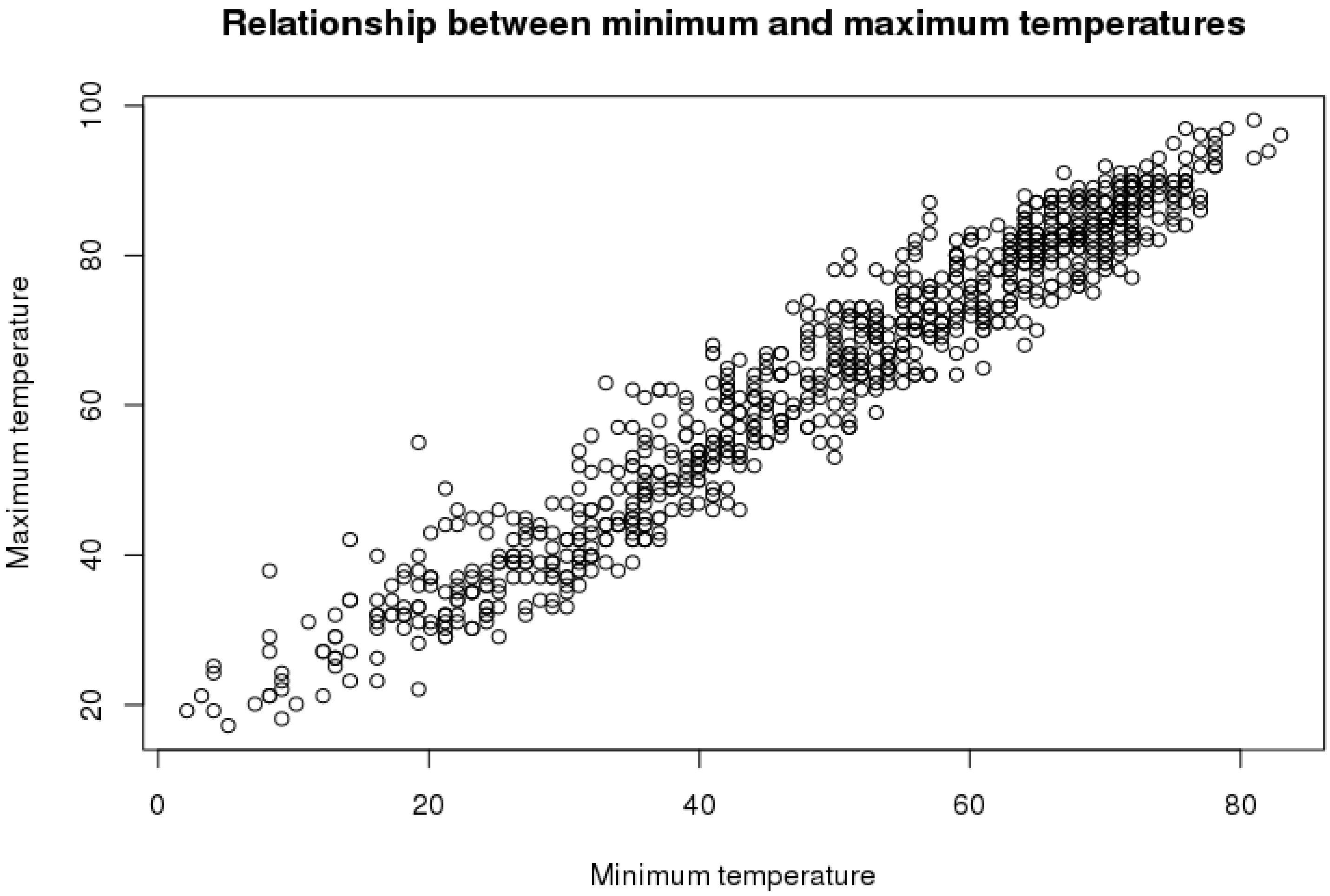
Scatter plots

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



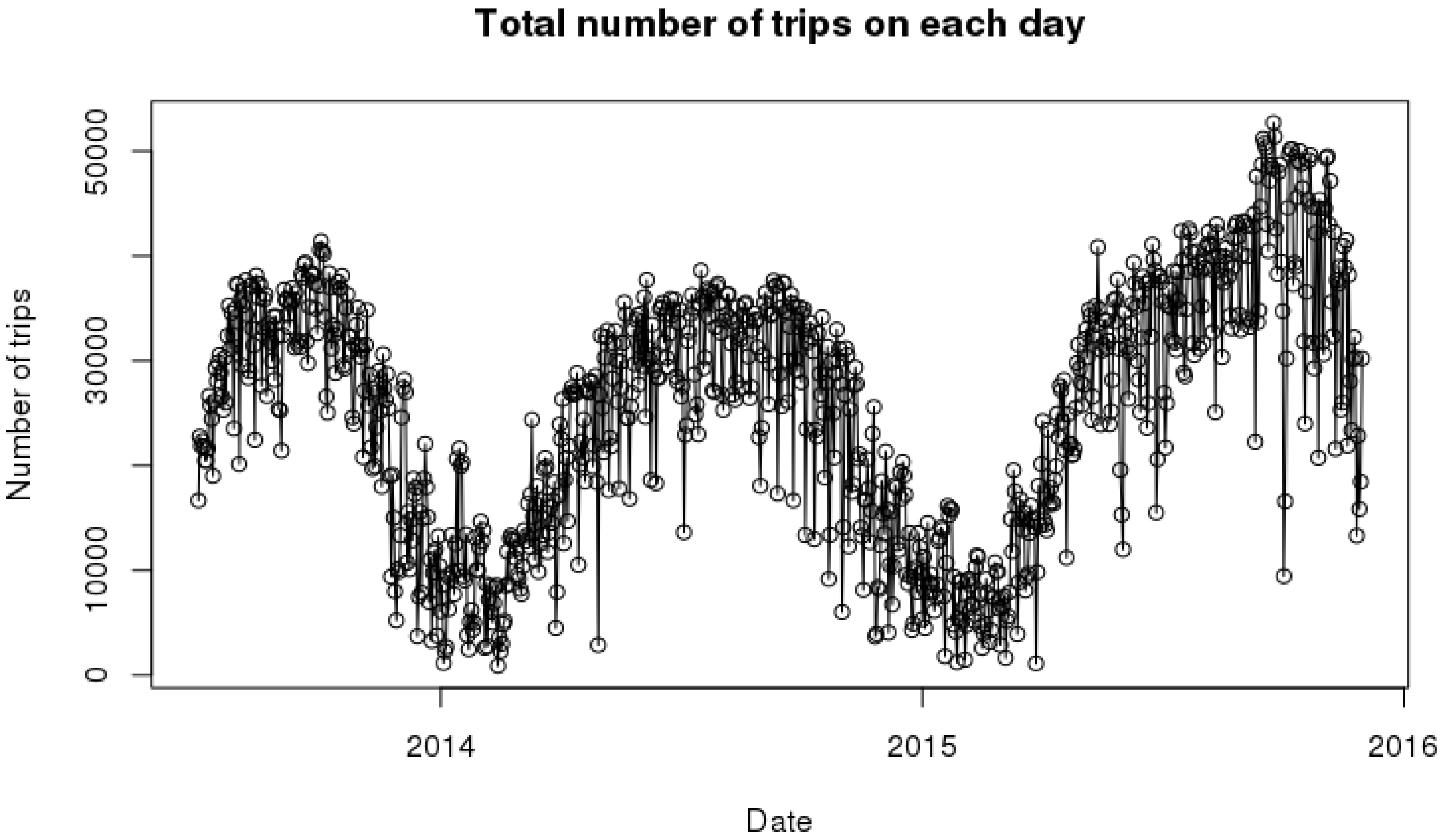
Plotting time series

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



Announcement: Homework 1

Homework 1: SDS230::download_homework(1)

Due on Gradescope by 11pm on Sunday September 11th

- 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!