431 Class 04

Thomas E. Love, Ph.D.

2022-09-08

Today's Agenda

- Look at the Five Questions posed during Class 03
- Make use of the data presented and cleaned in Class 03

From Class 03

Load packages and set theme

```
library(janitor)
library(patchwork)
library(tidyverse)

theme_set(theme_bw())
```

Read in data from `.csv` file

```
quicksur_raw <-
  read_csv("c04/data/quick_survey_2022.csv", show_col_types = FALSE) |>
  clean_names()
```

Manage the data into `qsdat`

Select variables

```
qsdat <- quicksur_raw |>
    select(student, year, english, smoke,
           pulse, height_in, haircut)
```

Change variable types

```
qsdat <- qsdat |>
    mutate(year = as_factor(year),
           smoke = as_factor(smoke),
           english = as_factor(english),
           student = as.character(student))
```

Where are we now?

```
1 summary(qsdat)
```

```
   student                year          english      smoke          pulse      
 Length:494           2020   : 67   n   :101   1   :456   Min.   : 30.00  
 Class :character     2016   : 64   y   :390   2   : 28   1st Qu.: 65.00  
 Mode  :character     2019   : 61   NA's:  3   3   :  8   Median : 72.00  
                      2021   : 58              NA's:  2   Mean   : 73.57  
                      2022   : 54                         3rd Qu.: 80.00  
                      2018   : 51                         Max.   :110.00  
                      (Other):139                         NA's   :75      
   height_in        haircut      
 Min.   :57.00   Min.   :  0.00  
 1st Qu.:64.00   1st Qu.: 14.00  
 Median :67.00   Median : 20.00  
 Mean   :67.33   Mean   : 30.17  
 3rd Qu.:70.00   3rd Qu.: 40.00  
 Max.   :77.50   Max.   :250.00  
 NA's   :7       NA's   :0       
```

Today's Questions

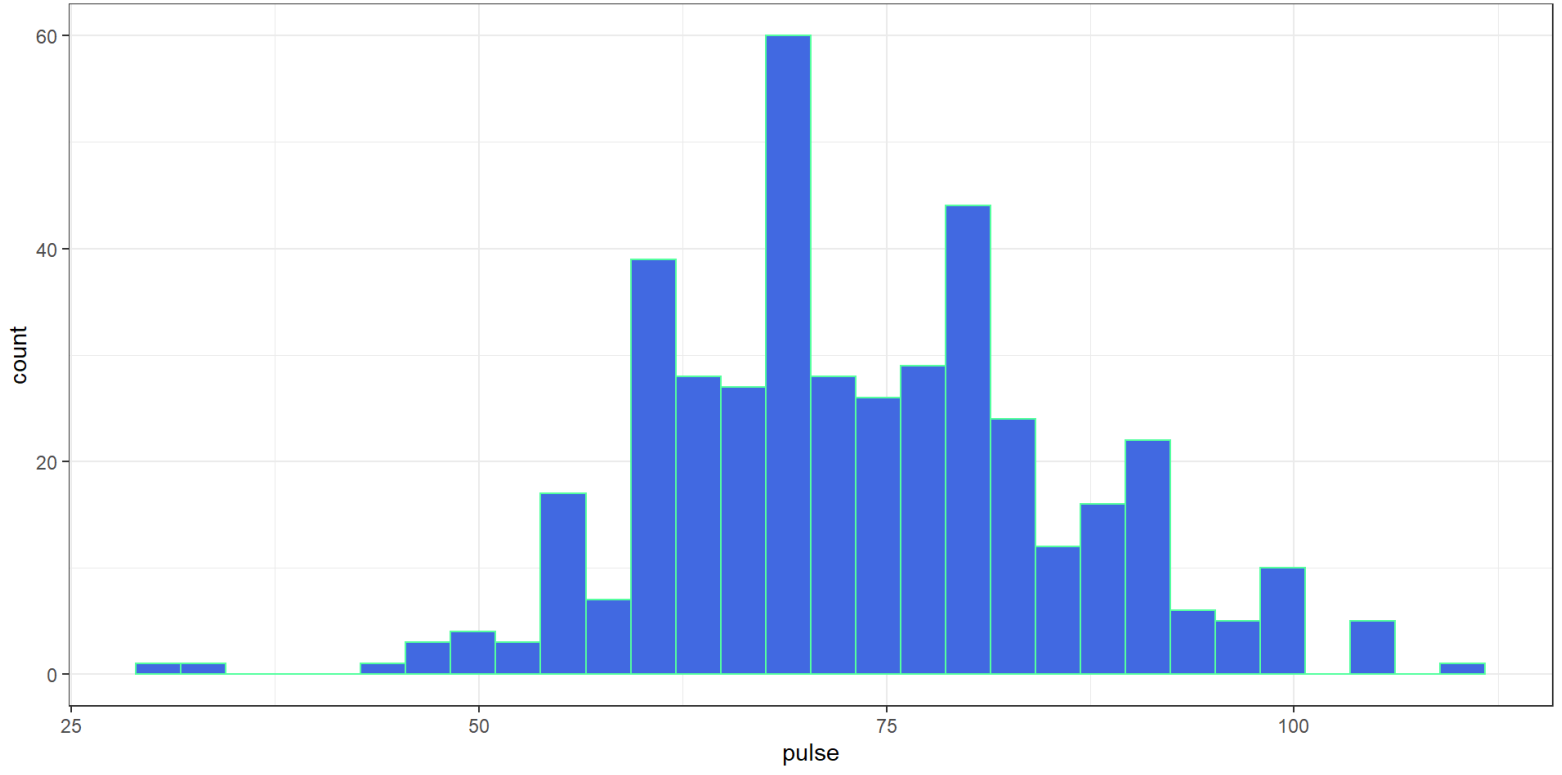
1. What is the distribution of pulse rates among students in 431 since 2014?
2. Does the distribution of student heights change materially over time?
3. Is a Normal distribution a good model for our data?
4. Do taller people appear to have paid less for their most recent haircut?
5. Do students have a more substantial tobacco history if they prefer to speak English or a language other than English?

Question 1 (Distribution of Student Pulse Rates)

Histogram, first try

- What is the distribution of student `pulse` rates?

```
1  ggplot(data = qsdat, aes(x = pulse)) +
2      geom_histogram(bins = 30, fill = "royalblue",  col = "seagreen1")
```



Describing the Pulse Rates

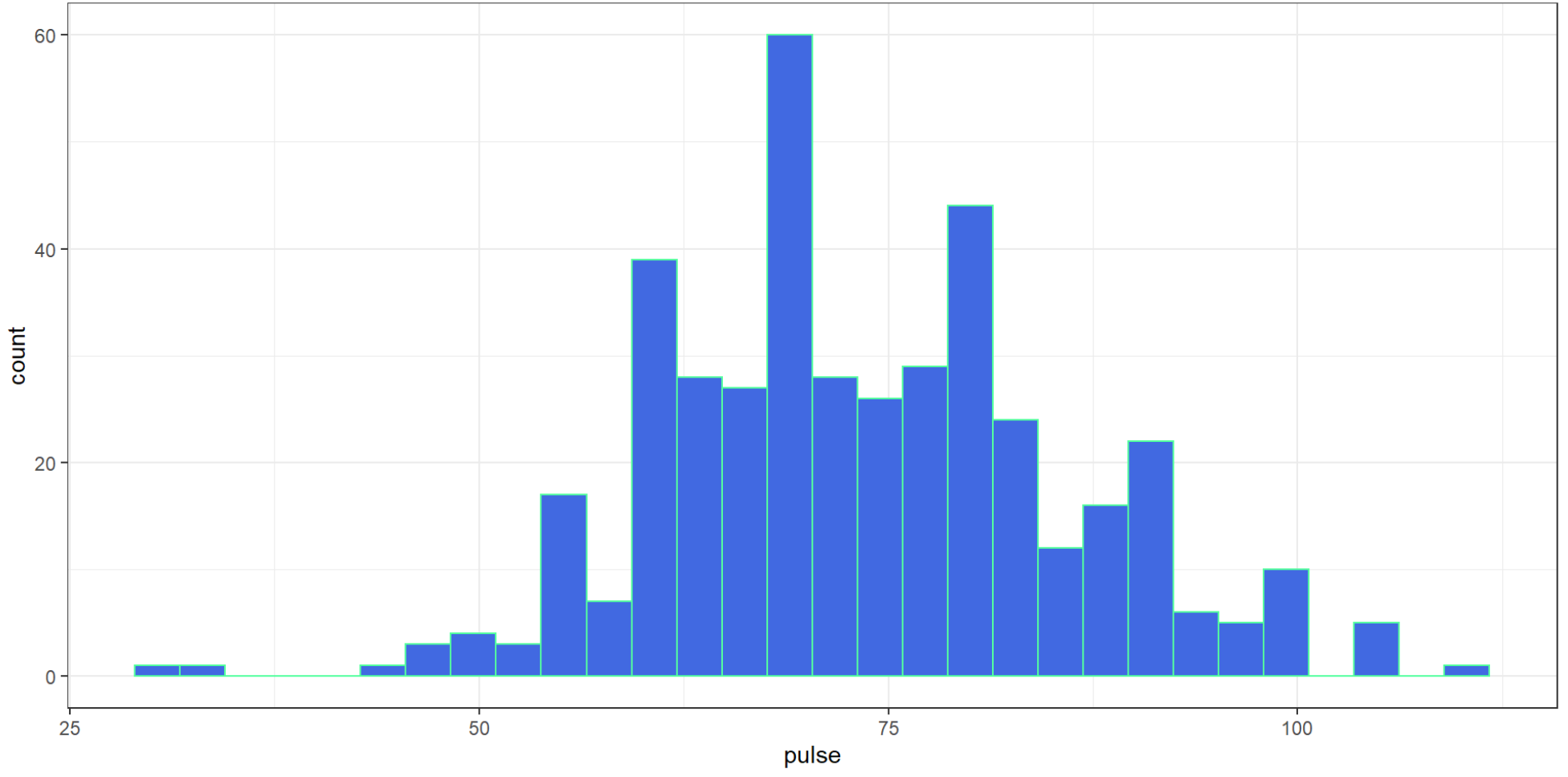
How might we describe this distribution?

- What is the center?
- How much of a range around that center do we see? How spread out are the data?
- What is the shape of this distribution?
  - Is it symmetric, or is it skewed to the left or to the right?

(Histogram is replotted on the next slide)

Histogram, first try again

```
1 ggplot(data = qsdat, aes(x = pulse)) +
2     geom_histogram(bins = 30, fill = "royalblue", col = "seagreen1")
```



Fundamental Numerical Summaries

```
1 qsdat |> select(pulse) |> summary()
```

```
     pulse      
 Min.   : 30.00  
 1st Qu.: 65.00  
 Median : 72.00  
 Mean   : 73.57  
 3rd Qu.: 80.00  
 Max.   :110.00  
 NA's   :75      
```

- How do the summary statistics help us describe the data?
- Do the values make sense to you?

```
1 mosaic::favstats(~ pulse, data = qsdat)
```

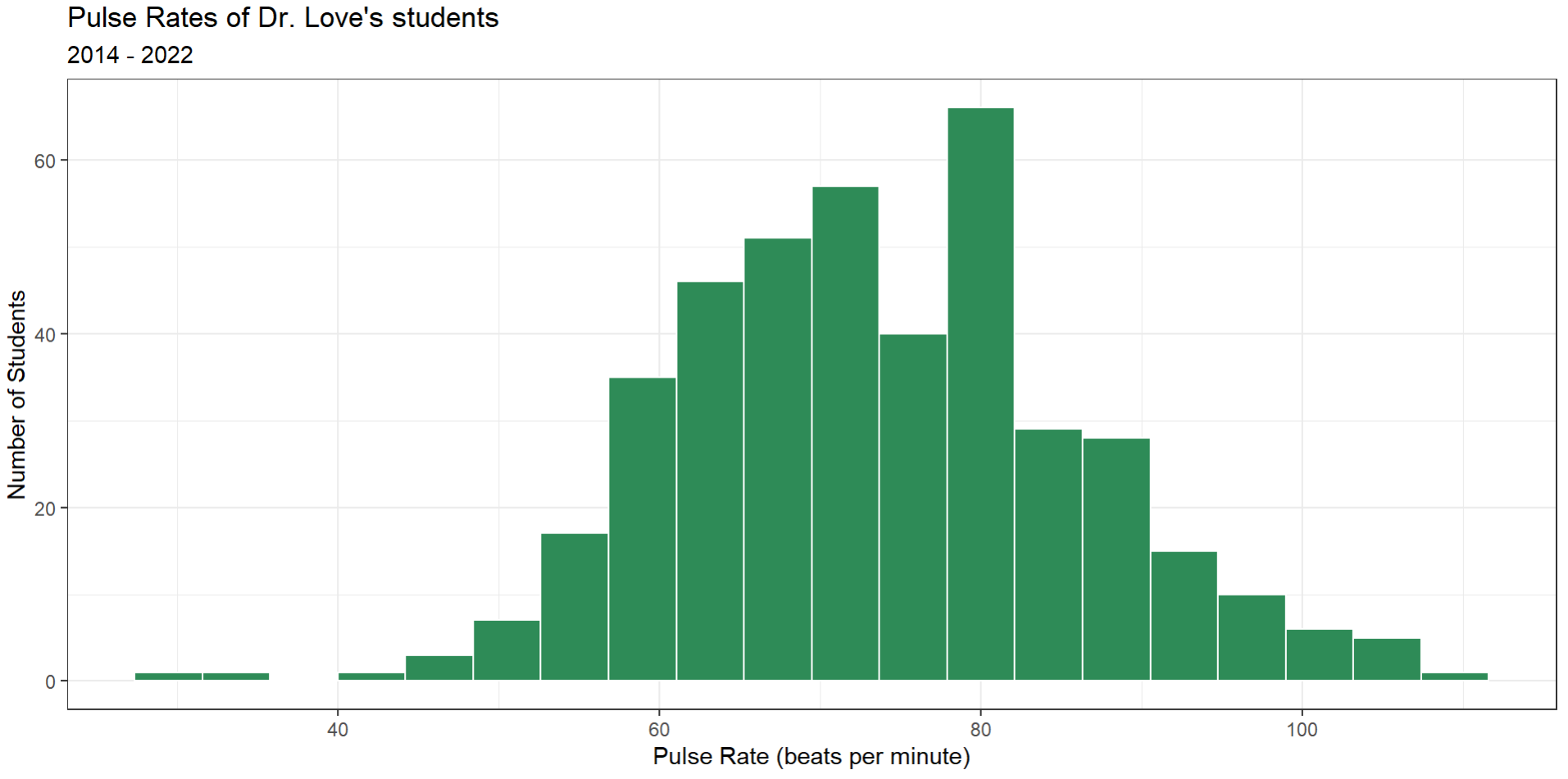
```
min Q1 median Q3 max     mean       sd   n missing
 30 65     72 80 110 73.57041 12.3539 419      75
```

Histogram, version 2

```
dat1 <- qsdat |>
  filter(complete.cases(pulse))

ggplot(data = dat1, aes(x = pulse)) +
    geom_histogram(fill = "seagreen", col = "white", bins = 20) +
    labs(title = "Pulse Rates of Dr. Love's students",
         subtitle = "2014 - 2022",
         y = "Number of Students",
         x = "Pulse Rate (beats per minute)")
```

- How did we deal with missing data?
- How did we add axis labels and titles to the plot?
- What is the distinction between `fill` and `col`?
- How many bins should we use?

Histogram, version 2



Question 2 (Student Heights over Time)

Yearly Five-Number Summaries

```
qsdat |>
    filter(complete.cases(height_in)) |>
    group_by(year) |>
    summarize(n = n(), min = min(height_in), q25 = quantile(height_in, 0.25
              median = median(height_in), q75 = quantile(height_in, 0.75),
              max = max(height_in))
```

- What should this produce? (Results on next slide)

Yearly Five-Number Summaries

```
# A tibble: 9 × 7
  year      n   min   q25 median   q75   max
  <fct> <int> <dbl> <dbl>  <dbl> <dbl> <dbl>
1 2014     40    60  64.8   68    71    73
2 2015     49    61  65     68    70    74
3 2016     64    60  64     67    70    76
4 2017     48    62  65     67    69    77
5 2018     51    60  63     66    70    73
6 2019     60    57  65     68    70    77.5
7 2020     66    59  63     66    69.8  76
8 2021     55    60  64.5   67.5  71    77.5
9 2022     54    59  66     68.5  70.4  76
```

- Does the distribution of heights change materially in 2014-2022?
- What are these summaries, specifically?

Five-Number Summary

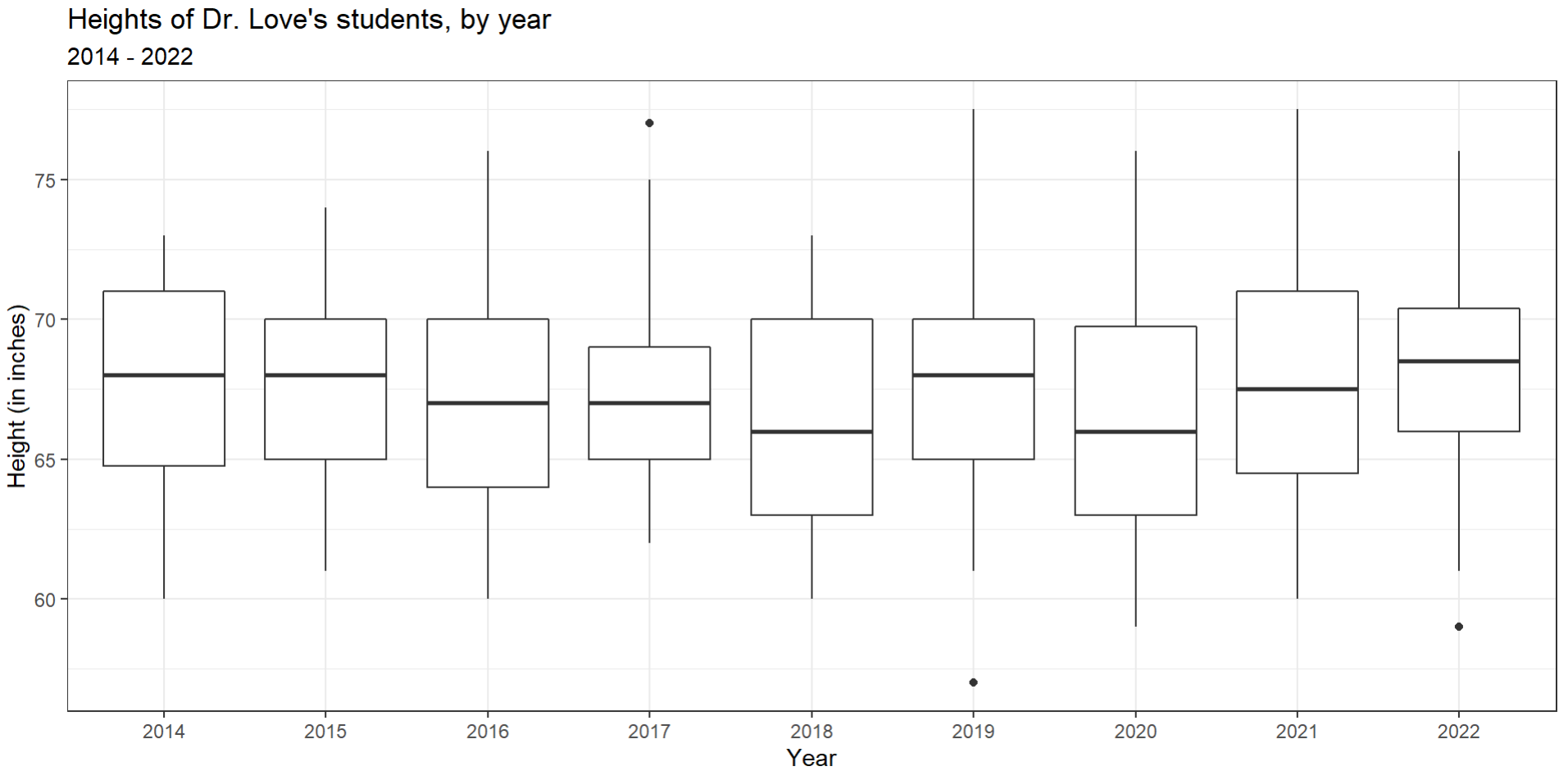
- Key summaries based on percentiles / quantiles
  - minimum = 0th, maximum = 100th, median = 50th
  - quartiles (25th, 50th and 75th percentiles)
  - Range is maximum - minimum
  - IQR (inter-quartile range) is 75th - 25th percentile
- These summaries are generally more resistant to outliers than mean, standard deviation
- Form the elements of a boxplot (box-and-whisker plot)

Comparison Boxplot of Heights by Year

```
dat2 <- qsdat |>
    filter(complete.cases(height_in))

ggplot(data = dat2, aes(x = year, y = height_in)) +
    geom_boxplot() +
    labs(title = "Heights of Dr. Love's students, by year",
         subtitle = "2014 - 2022", x = "Year", y = "Height (in inches)")
```

- How did we deal with missing data here?

Comparison Boxplot of Heights by Year



Thinking about the Boxplot

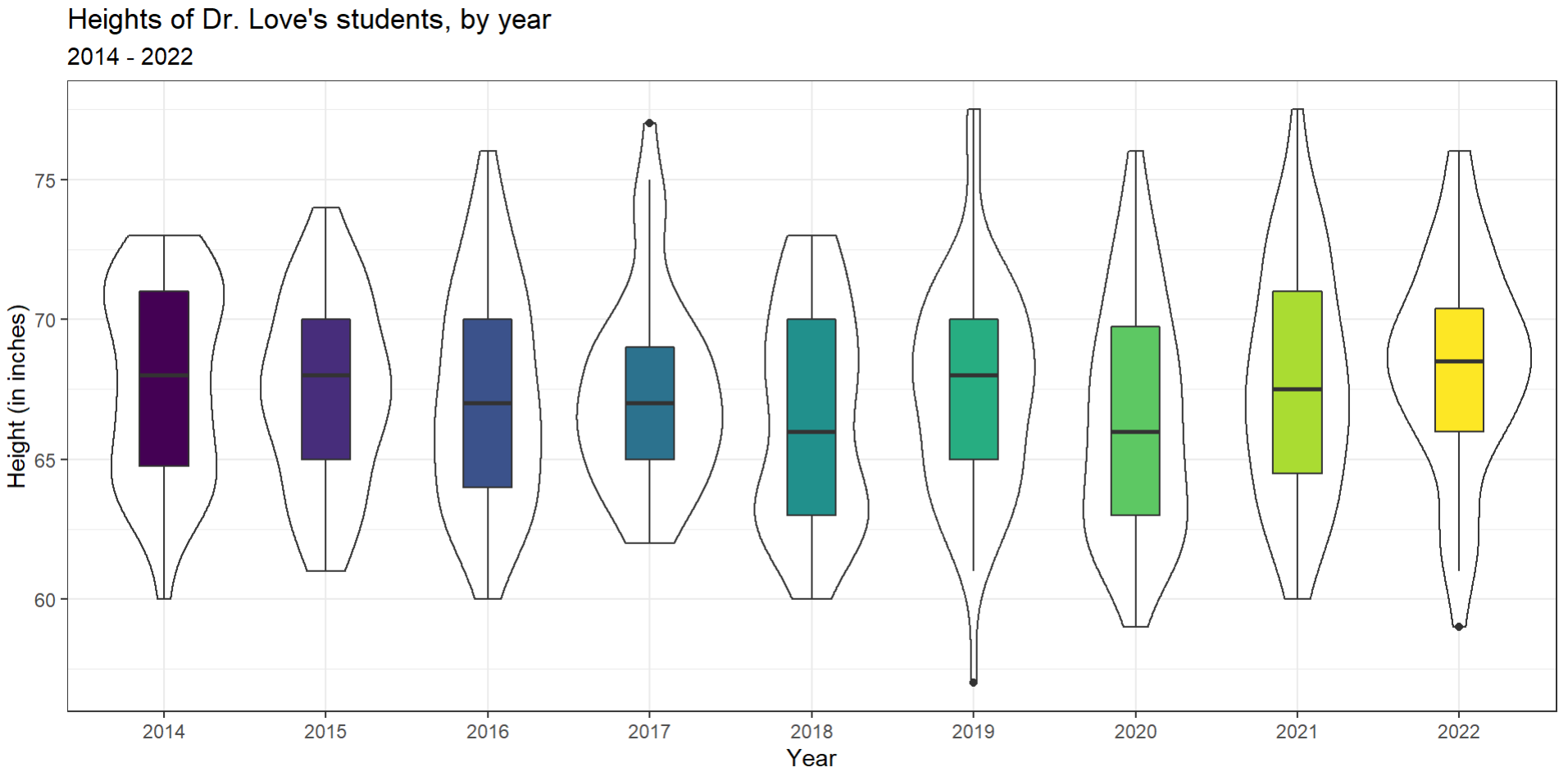
- Box covers the middle half of the data (25th and 75th percentiles), and the solid line indicates the median
- Whiskers extend from the quartiles to the most extreme values that are not judged by **Tukey's** "fences" method to be candidate outliers
  - Fences are drawn at 25th percentile - 1.5 IQR and 75th percentile + 1.5 IQR
- Are any values candidate outliers by this method? For which years?
- Was it important to change `year` to a factor earlier?

Adding a Violin to the Boxplot

- When we'd like to better understand the shape of a distribution, we can amplify the boxplot.

```
dat2 <- qsdat |>
    filter(complete.cases(height_in))

ggplot(data = dat2, aes(x = year, y = height_in)) +
    geom_violin() +
    geom_boxplot(aes(fill = year), width = 0.3) +
    guides(fill = "none") +
    scale_fill_viridis_d() +
    labs(title = "Heights of Dr. Love's students, by year",
         subtitle = "2014 - 2022", x = "Year", y = "Height (in inches)")
```

Adding a Violin to the Boxplot



Thinking About our Boxplot with Violin

- How did we change the boxplot when we added the violin?
- What would happen if we added the boxplot first and the violin second?
- What does `guides(fill = "none")` do?
- What does `scale_fill_viridis_d()` do?

Table of Means and Standard Deviations

```
qsdat |>
    filter(complete.cases(height_in)) |>
    group_by(year) |>
    summarize(n = n(), mean = mean(height_in), sd = sd(height_in))
```

```
# A tibble: 9 × 4
  year      n  mean    sd
  <fct> <int> <dbl> <dbl>
1 2014     40  67.8  3.46
2 2015     49  67.3  3.32
3 2016     64  67.2  3.86
4 2017     48  67.4  3.46
5 2018     51  66.5  3.81
6 2019     60  67.4  3.83
7 2020     66  66.4  4.09
8 2021     55  67.8  4.13
9 2022     54  68.4  3.74
```

So, what do we think?

Are the distributions of student height very different from year to year?

- What output that I've provided here can help answer this question?
- What other things would you like to see?

Question 3
Can we assume that the Mean and SD are sensible summaries?

A Normal distribution (bell-shaped curve)

This is a Normal (or Gaussian) distribution with mean 150 and standard deviation 30.

Summarizing Quantitative Data

If the data followed a Normal model,

- we would be justified in using the sample **mean** to describe the center, and
- in using the sample **standard deviation** to describe the spread (variation.)

But it is often the case that these measures aren't robust enough, because the data show meaningful skew (asymmetry), or the data have lighter or heavier tails than a Normal model would predict.

The Empirical Rule for Approximately Normal Distributions

If the data followed a Normal distribution,

- approximately 68% of the data would be within 1 SD of the mean,
- approximately 95% of the data would be within 2 SD of the mean, while
- essentially all (99.7%) of the data would be within 3 SD of the mean.

Empirical Rule & 2022 Student Heights

In 2022, we had 54 students whose `height_in` was available, with mean 68.4 inches (173.7 cm) and standard deviation 3.7 inches (9.4 cm).

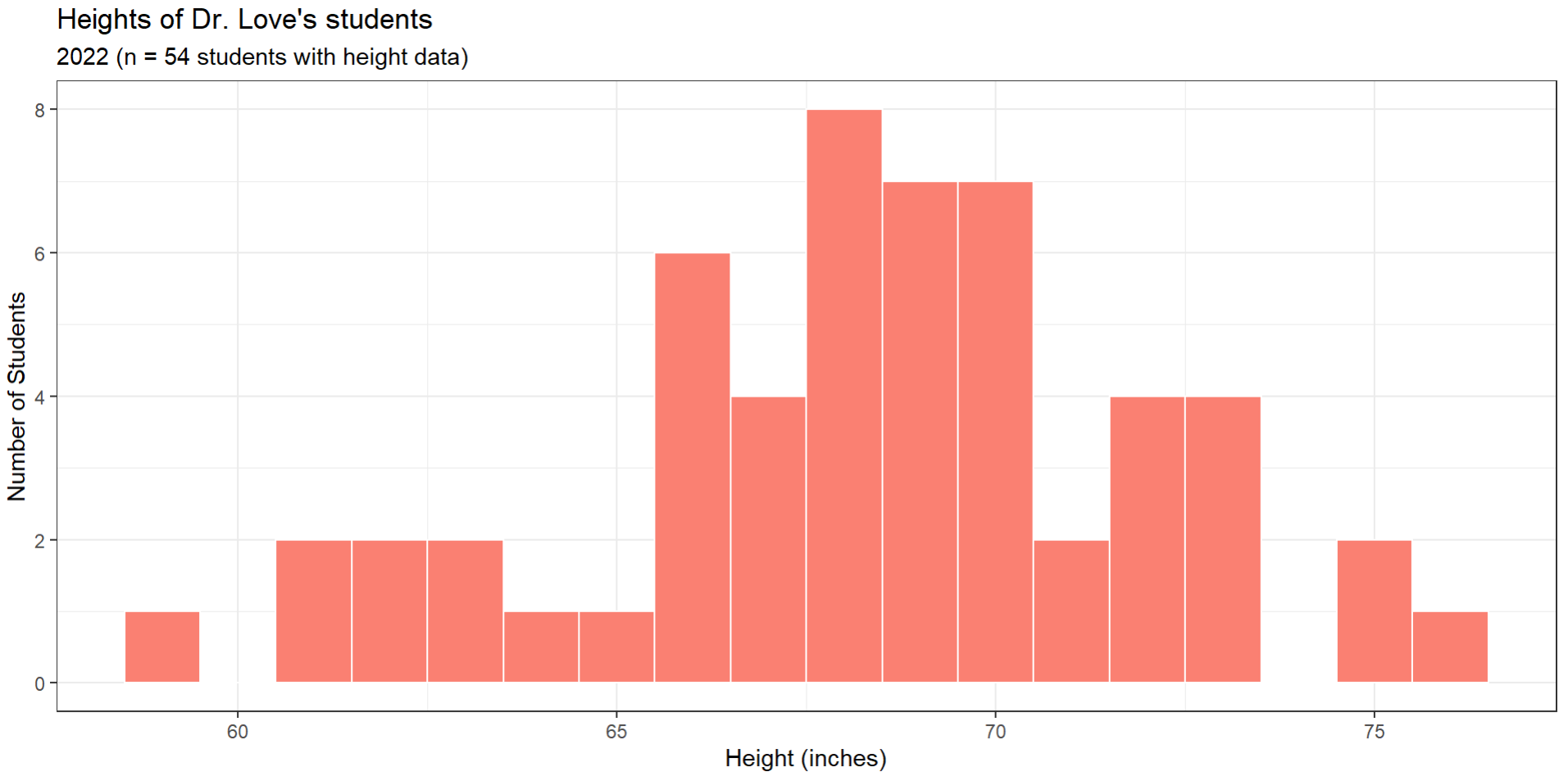
What do the histogram (next slide) and boxplot (seen earlier) suggest about whether a Normal model with this mean and standard deviation would hold well for these 54 student heights?

Histogram of 2022 Student Heights

```
dat3 <- qsdat |>
  filter(complete.cases(height_in)) |>
  filter(year == "2022")

ggplot(data = dat3, aes(x = height_in)) +
    geom_histogram(fill = "salmon", col = "white", binwidth = 1) +
    labs(title = "Heights of Dr. Love's students",
         subtitle = "2022 (n = 54 students with height data)",
         y = "Number of Students", x = "Height (inches)")
```

- How did we use the two `filter()` statements?
- Why might I have changed from specifying `bins` to `binwidth` here?

Histogram of 2022 Student Heights



Checking the 1-SD Empirical Rule

- Of the 54 students in 2022 with heights, how many were within 1 SD of the mean?
  - Mean = 68.4, SD = 3.7.
  - 68.4 - 3.7 = 64.7 inches and 68.4 + 3.7 = 72.1 inches

```
qsdat |> filter(complete.cases(height_in)) |>
    filter(year == "2022") |>
    count(height_in >= 64.7 & height_in <= 72.1)
```

```
# A tibble: 2 × 2
  `height_in >= 64.7 & height_in <= 72.1`     n
  <lgl>                                   <int>
1 FALSE                                      15
2 TRUE                                       39
```

```
39/(39+15)
```

```
[1] 0.7222222
```

2-SD Empirical Rule

- How many of the 54 `height_in` values gathered in 2022 were between 68.4 - 2(3.7) = 61.0 and 68.4 + 2(3.7) = 75.8 inches?

```
qsdat |> filter(complete.cases(height_in)) |>
    filter(year == "2022") |>
    count(height_in >= 61.0 & height_in <= 75.8)
```

```
# A tibble: 2 × 2
  `height_in >= 61 & height_in <= 75.8`     n
  <lgl>                                 <int>
1 FALSE                                     2
2 TRUE                                     52
```

```
52/(52+2)
```

```
[1] 0.962963
```

3-SD Empirical Rule

- How many of the 54 `height_in` values gathered in 2022 were between 68.4 - 3(3.7) = 57.3 and 68.4 + 3(3.7) = 79.5 inches?

```
qsdat |> filter(complete.cases(height_in)) |>
    filter(year == "2022") |>
    count(height_in >= 57.3 & height_in <= 79.5)
```

```
# A tibble: 1 × 2
  `height_in >= 57.3 & height_in <= 79.5`     n
  <lgl>                                   <int>
1 TRUE                                       54
```

```
54/(54+0)
```

```
[1] 1
```

Empirical Rule Table for 2022 data

- $\bar{x}$ = sample mean, s = sample SD
- For `height_in`: n = 54 with data, $\bar{x}$ = 68.4, s = 3.7$
- For `pulse`: n = 52 with data, $\bar{x}$ = 75.4, s = 11.2$

Range	"Normal"	`height_in`	`pulse`
$\bar{x} \pm s$	~68%	$\frac{39}{54}$ = 72.2%	$\frac{43}{52}$ = 82.7%

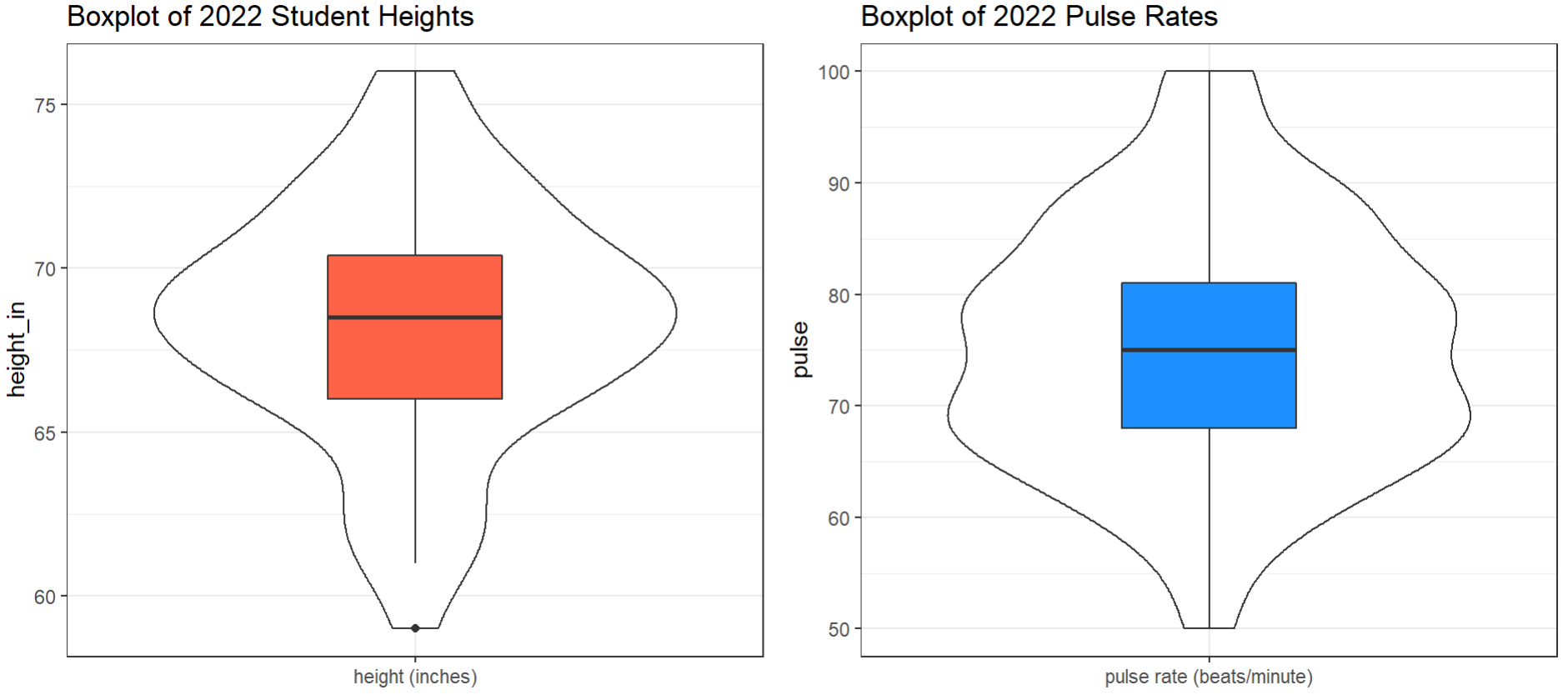
Boxplots of Height and of Pulse Rate

```
dat4 <- qsdat |> filter(complete.cases(height_in), year == "2022")

p4 <- ggplot(data = dat4, aes(x = "height (inches)", y = height_in)) +
  geom_violin() + geom_boxplot(width = 0.3, fill = "tomato") +
  labs(title = "Boxplot of 2022 Student Heights", x = "")

dat5 <- qsdat |> filter(complete.cases(pulse), year == "2022")

p5 <- ggplot(data = dat5, aes(x = "pulse rate (beats/minute)", y = pulse))
  geom_violin() + geom_boxplot(width = 0.3, fill = "dodgerblue") +
  labs(title = "Boxplot of 2022 Pulse Rates", x = "")

p4 + p5 +
  plot_annotation(title = "2022 Quick Survey Data")
```

- What is `width = 0.3` doing? How about the `x` options?
- What am I doing with `p3 + p4 + plot_annotation`?
- What should this look like?

Boxplots of Height and of Pulse Rate

2022 Quick Survey Data



Normality and Mean/SD as summaries

If the data are approximately Normally distributed (like `height_in` and `pulse`) we can safely use the sample mean and standard deviation as summaries. If not "Normal", then …

- The median is a more robust summary of the center.
- For spread, we often use the 25th and 75th percentiles.

```
1 dat3 <- qsdat |> filter(year == "2022")
2 mosaic::favstats(~ height_in, data = dat3)
```

```
min Q1 median     Q3 max     mean       sd  n missing
 59 66   68.5 70.375  76 68.35185 3.742451 54       0
```

```
1 mosaic::favstats(~ pulse, data = dat3)
```

```
min Q1 median Q3 max     mean       sd  n missing
 50 68     75 81 100 75.40385 11.17466 52       2
```

A new quantitative variable

Let's look at haircut prices, across all years.

```
1 mosaic::favstats(~ haircut, data = qsdat)
```

```
 min Q1 median Q3 max     mean       sd   n missing
   0 14     20 40 250 30.17214 31.4079 485       9
```

Does it seem like the Normal model will be a good fit for these prices?

- Why or why not?
- What more information do you need to make a decision?

2022 Haircut Prices

Unsorted | Sorted | Counts

```
1  qsdat |> filter(year == "2022") |>
2    select(haircut) |>
3    as.vector() ## just to print it here horizontally
```

```
$haircut
 [1]   2  50   0  38  25  15   0  32  60  40  45  52  52   0  30  15  30  75
20
[20]   4  45  20  35   0  50  25  40 240  30   6  45  25   2  30  25  20 200
15
[39]  35   0  20  80  20  30  80  10  50  30  60  20  20  30   5  30
```

2022 Haircut Prices, tabulated

```
1 qsdat |> filter(year == "2022") |> tabyl(haircut) |> adorn_pct_formatting()
```

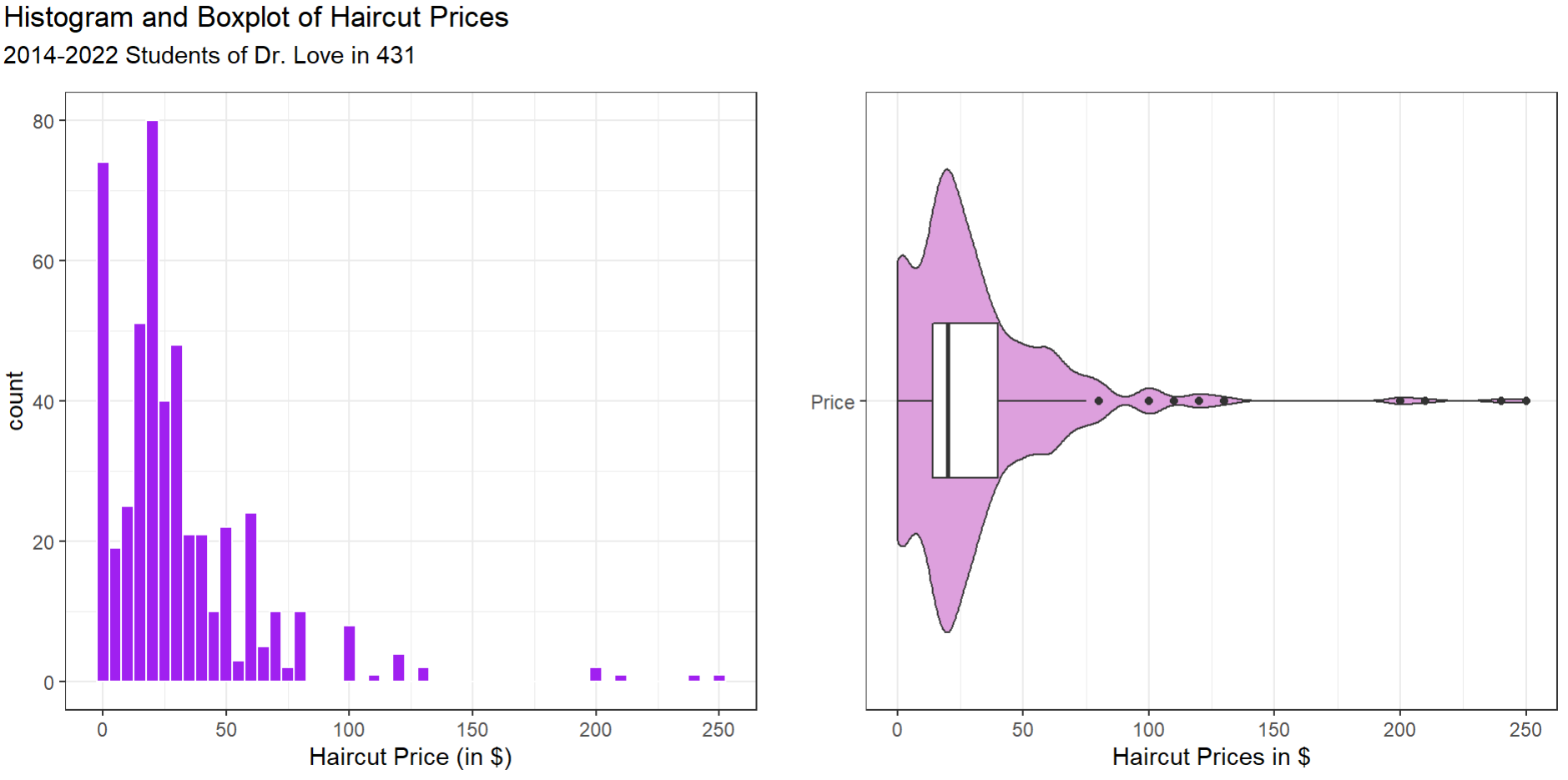
haircut	n	percent
0	5	9.3%
2	2	3.7%
4	1	1.9%
5	1	1.9%
6	1	1.9%
10	1	1.9%
15	3	5.6%
20	7	13.0%
25	4	7.4%
30	8	14.8%
32	1	1.9%
35	2	3.7%
38	1	1.9%
40	2	3.7%
45	2	5.6%

Normality of Haircut prices?

```
dat6 <- qsdat |> filter(complete.cases(haircut))

p6a <- ggplot(data = dat6, aes(x = haircut)) +
  geom_histogram(binwidth = 5, fill = "purple", col = "white") +
  labs(x = "Haircut Price (in $)")

p6b <- ggplot(data = dat6, aes(x = haircut, y = "Price")) +
  geom_violin(fill = "plum") + geom_boxplot(width = 0.3) +
  labs(y = "", x = "Haircut Prices in $")

p6a + p6b +
  plot_annotation(
    title = "Histogram and Boxplot of Haircut Prices",
    subtitle = "2014-2022 Students of Dr. Love in 431")
```

- Do you think that the distribution of these prices follows a Normal model?

Normality of Haircut prices?



Stem-and-Leaf of Haircut Prices

```
1 stem(qsdat$haircut, scale = 2) # scale makes plot twice as long as default
```

```
The decimal point is 1 digit(s) to the right of the |

 0 | 00000000000000000000000000000000000000000000000000000000000000000000+19
 1 | 00000002222222222334444444555555555555555555555555555555556667
 2 | 00000000000000000000000000000000000000000000000000000000000012+34
 3 | 00000000000000000000000000000000122222555555555555555555688
 4 | 00000000000000000035555555558
 5 | 00000000000000000022555
 6 | 00000000000000000000055555
 7 | 000000002255
 8 | 0000000000
 9 |
10 | 00000000
11 | 0
```

- Note this is *not* a `ggplot` so it works differently than most plots we will make this term.

Empirical Rule Table for Haircut Prices

Let's look across all years, as well as just in 2022

```
1 mosaic::favstats(~ haircut, data = qsdat)
```

```
min Q1 median Q3 max     mean       sd   n missing
  0 14     20 40 250 30.17214 31.4079 485       9
```

```
1 mosaic::favstats(~ haircut, data = qsdat |> filter(year == "2022"))
```

```
min    Q1 median Q3 max     mean       sd  n missing
  0 16.25     30 45 240 36.25926 41.81955 54       0
```

Range	"Normal"	2014-2022	2022
\(\bar{x} \pm s\)	~68%	\(\frac{438}{485}\) = 90.3%	\(\frac{50}{54}\) = 92.6%

Range	“Normal”	2014-2022	2022
$\bar{x} \pm 2\times s$	~95%	$\frac{465}{485}$ = 95.6%	$\frac{52}{54}$ = 96.3%
$\bar{x} \pm 3\times s$	~99.7%	$\frac{478}{485}$ = 98.6%	$\frac{52}{54}$ = 96.3%

How did I calculate those fractions?

```
# haircut price mean = 30.17 and sd = 31.41 across 2014-2022

qsdat |> count(haircut >= 30.17 - 31.41 & haircut <= 30.17 + 31.41)
qsdat |> count(haircut >= 30.17 - 2*31.41 & haircut <= 30.17 + 2*31.41)
qsdat |> count(haircut >= 30.17 - 3*31.41 & haircut <= 30.17 + 3*31.41)

# haircut price mean = 36.26 and sd = 41.82 in 2022 alone

qsdat |> filter(year == "2022") |>
  count(haircut >= 36.26 - 41.82 & haircut <= 36.26 + 41.82)
qsdat |> filter(year == "2022") |>
  count(haircut >= 36.26 - 2*41.82 & haircut <= 36.26 + 2*41.82)
qsdat |> filter(year == "2022") |>
  count(haircut >= 36.26 - 3*41.82 & haircut <= 36.26 + 3*41.82)
```

Question 4 (Heights and Haircut Prices)

Do tall people pay less for haircuts?

Why might we think that they do, before we see the data?

- Convert our student heights from inches to centimeters…

```
1 qsdat <- qsdat |> mutate(height_cm = height_in * 2.54)
2
3 qsdat |> select(student, height_in, height_cm) |> head()
```

```
# A tibble: 6 × 3
  student height_in height_cm
  <chr>       <dbl>     <dbl>
1 202201       69.5      177.
2 202202       63        160.
3 202203       73        185.
4 202204       70        178.
5 202205       59        150.
6 202206       68        173.
```

Initial Numerical Summaries

```
1 qsdat |> filter(complete.cases(haircut, height_cm)) |>
2   summarize(n = n(), median(haircut), median(height_cm), median(height_in))
```

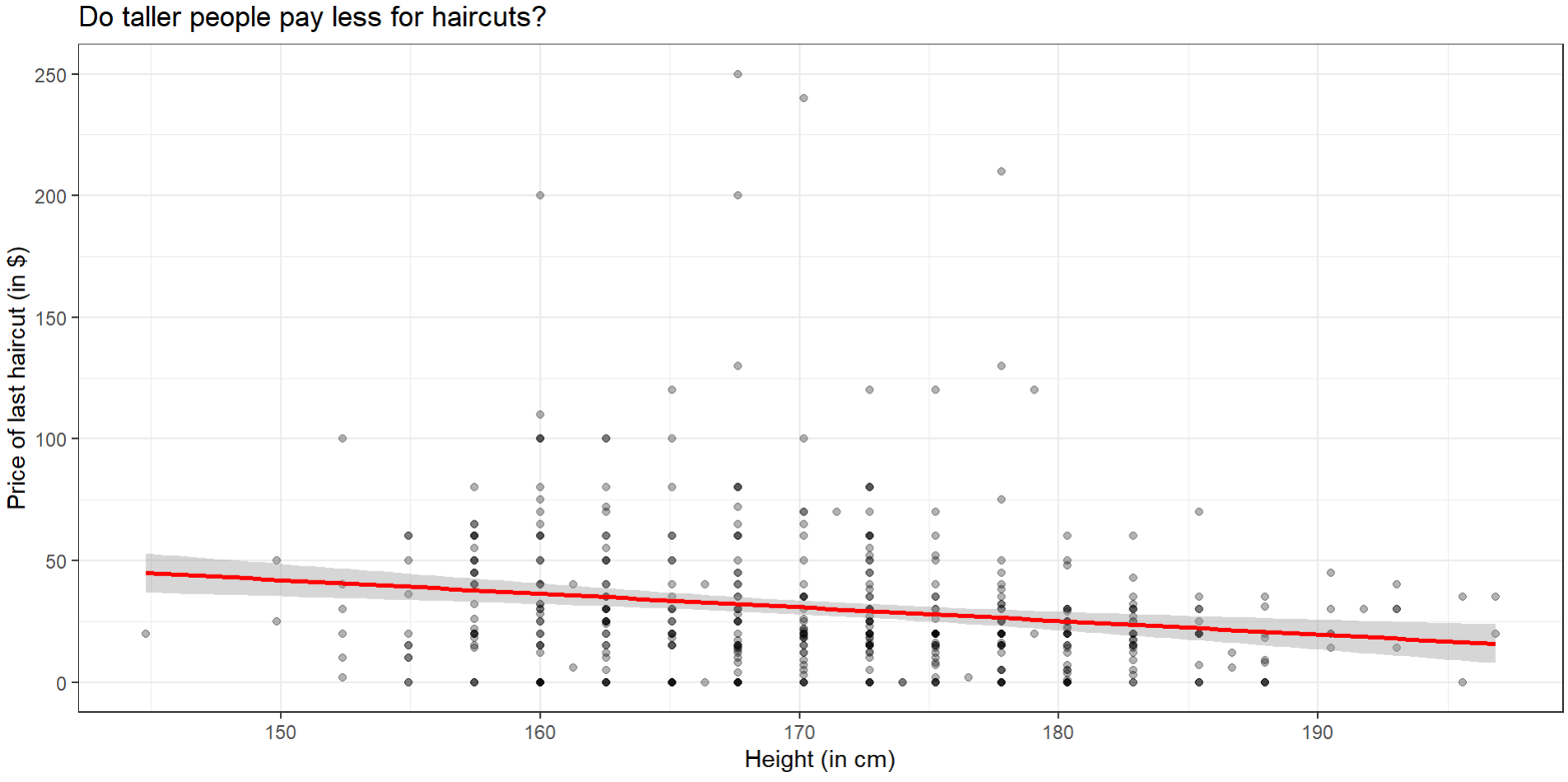
```
# A tibble: 1 × 4
      n `median(haircut)` `median(height_cm)` `median(height_in)`
  <int>             <dbl>               <dbl>               <dbl>
1   483                20                170.                  67
```

A First Scatterplot

- We'll include the straight line from a linear model, in red.

```
dat7 <- qsdat |> filter(complete.cases(height_cm, haircut))

ggplot(dat7, aes(x = height_cm, y = haircut)) +
    geom_point(alpha = 0.3) +
    geom_smooth(method = "lm", col = "red",
                formula = y ~ x, se = TRUE) +
    labs(x = "Height (in cm)",
         y = "Price of last haircut (in $)",
         title = "Do taller people pay less for haircuts?")
```

A First Scatterplot



What is the (Pearson) correlation of height and haircut price?

```
1 dat7 <- qsdat |> filter(complete.cases(height_cm, haircut))
2
3 dat7 |>
4     select(height_in, height_cm, haircut) |>
5     cor()
```

```
           height_in  height_cm    haircut
height_in  1.0000000  1.0000000 -0.1708551
height_cm  1.0000000  1.0000000 -0.1708551
haircut   -0.1708551 -0.1708551  1.0000000
```

What is the straight line regression model?

```
1 dat7 <- qsdat |> filter(complete.cases(height_cm, haircut))
2
3 mod1 <- lm(haircut ~ height_cm, data = dat7)
4
5 mod1
```

```
Call:
lm(formula = haircut ~ height_cm, data = dat7)

Coefficients:
(Intercept)    height_cm
   125.9519      -0.5597
```

Summarizing our model mod1

```
1 summary(mod1)
```

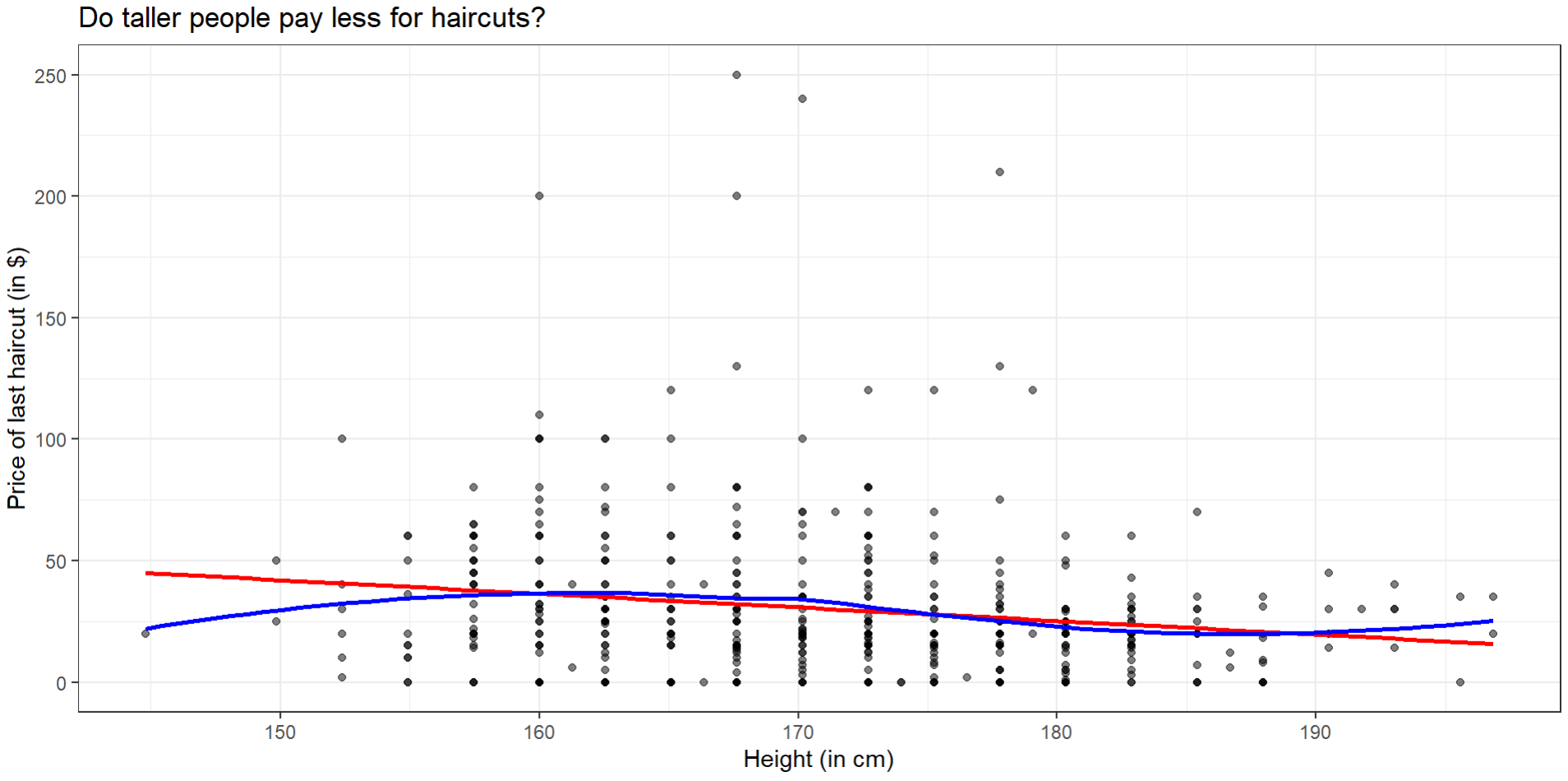
```
Call:
lm(formula = haircut ~ height_cm, data = dat7)

Residuals:
    Min      1Q  Median      3Q     Max 
-39.233 -18.124  -6.095   8.165 217.876 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 125.9519    25.2161   4.995 8.25e-07 ***
height_cm    -0.5597     0.1472  -3.803 0.000161 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 31.04 on 401 degrees of freedom
```

Compare `lm` fit to `loess` smooth curve?

```
dat7 <- qsdat |> filter(complete.cases(height_cm, haircut))

ggplot(dat7, aes(x = height_cm, y = haircut)) +
    geom_point(alpha = 0.5) +
    geom_smooth(method = "lm", col = "red",
                formula = y ~ x, se = FALSE) +
    geom_smooth(method = "loess", col = "blue",
                formula = y ~ x, se = FALSE) +
    labs(x = "Height (in cm)",
         y = "Price of last haircut (in $)",
         title = "Do taller people pay less for haircuts?")
```

- Does a linear model appear to fit these data well?
- Do taller people pay less for their haircuts?

Compare `lm` fit to `loess` smooth curve?



Question 5 (Tobacco and Language Preference)

Restrict ourselves to 2022 data

- Do students in the 2022 class have a more substantial history of tobacco use if they prefer to speak a language other than English?

```
dat9 <- qsdat |>
    filter(year == "2022") |>
    select(student, year, english, smoke)
```

```
summary(dat9)
```

```
  student                year     english smoke
Length:54           2022   :54   n:16    1:47
Class :character    2014   : 0   y:38    2: 6
Mode  :character    2015   : 0           3: 1
                    2016   : 0
                    2017   : 0
                    2018   : 0
                    (Other): 0
```

No missing data.

Tabulating the categorical variables individually

```
1 dat9 |> tabyl(english)
```

```
english  n   percent
      n 16 0.2962963
      y 38 0.7037037
```

```
1 dat9 |> tabyl(smoke) |> adorn_pct_formatting()
```

```
smoke  n percent
    1 47   87.0%
    2  6   11.1%
    3  1    1.9%
```

- What does `adorn_pct_formatting()` do?

Cross-Classification (2 rows \(\times\) 3 columns)

```
1 dat9 |> tabyl(english, smoke)
```

```
english  1 2 3
      n 15 0 1
      y 32 6 0
```

Recode the `smoke` levels to more meaningful names in `tobacco`

```
dat9 <- dat9 |>
    mutate(tobacco = fct_recode(smoke,
                "Never" = "1", "Quit" = "2", "Current" = "3"))
```

Check our work?

```
dat9 |> count(smoke, tobacco)
```

```
# A tibble: 3 × 3
  smoke tobacco     n
  <fct> <fct>   <int>
1 1     Never      47
2 2     Quit        6
3 3     Current     1
```

- Everyone with `smoke` = 1 has `tobacco` as Never, etc.

Restate the cross-tabulation

Now we'll use this new variable, and this time, add row and column totals.

```
1 dat9 |> tabyl(english, tobacco) |>
2     adorn_totals(where = c("row", "col"))
```

```
english Never Quit Current Total
      n    15    0       1    16
      y    32    6       0    38
  Total    47    6       1    54
```

- What can we conclude about this association?

How about in 2014-2022?

```
dat8 <- qsdat |>
  filter(complete.cases(english, smoke)) |>
  mutate(tobacco = fct_recode(smoke,
            "Never" = "1", "Quit" = "2", "Current" = "3"))

dat8 |>
  tabyl(english, tobacco) |>
  adorn_totals(where = c("row", "col"))
```

english	Never	Quit	Current	Total
n	95	2	4	101
y	359	26	4	389
Total	454	28	8	490

- Now, what is your conclusion?

Next Time

Analyzing a (small) health dataset

Cleaning up the temporary objects

```
rm(mod1,
   p4, p5, p6a, p6b,
   dat1, dat2, dat3, dat4, dat5, dat6, dat7, dat8, dat9
   )

## this just leaves
## qsdat and quicksur_raw in my Global Environment
```

Session Information

Don't forget to close your file with the session information.

```
1 sessionInfo()
```

```
R version 4.2.1 (2022-06-23 ucrt)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 10 x64 (build 22000)

Matrix products: default

locale:
[1] LC_COLLATE=English_United States.utf8
[2] LC_CTYPE=English_United States.utf8
[3] LC_MONETARY=English_United States.utf8
[4] LC_NUMERIC=C
[5] LC_TIME=English_United States.utf8

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base
```