431 Class 03

Thomas E. Love

2019-09-03

Today's Agenda

Using R, RStudio and R Markdown and the 431 RStudio Cloud

Contact us at `431-help@case.edu`

Our web site: https://github.com/THOMASELOVE/2019-431

RStudio Cloud In-Class Early Project

We assume you were able to follow the software installation instructions.

If so, you'd want to:

1. Get data from our site to a new directory on your machine.
2. Open RStudio and start a new Project, in the new directory.
3. Open and set up an R Markdown file to do the work.

But, perhaps you haven't gotten to that yet. So we have RStudio Cloud.

Link to join is: http://bit.ly/431-2019-join-cloud

First Step: Load the Packages You Need

```r
library(tidyverse)
```

```
-- Attaching packages ---------------------------------------

v ggplot2 3.2.0     v purrr   0.3.2
v tibble  2.1.3     v dplyr   0.8.3
v tidyr   0.8.3     v stringr 1.4.0
v readr   1.3.1     v forcats 0.4.0

-- Conflicts ------------------------------------------------
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
```

Analyzing the Index Card Guesses of My Age

61 students turned in an index card, meant to contain both a first and a second guess of my age.

For the slides, I have this information in a subfolder called data in my R Project.

```
love_2019 <- read_csv("data/love-age-guess-2019.csv")
```

```
Parsed with column specification:
cols(
  subject = col_character(),
  age1 = col_double(),
  age2 = col_double()
)
```

The love_2019 tibble

```
love_2019
```

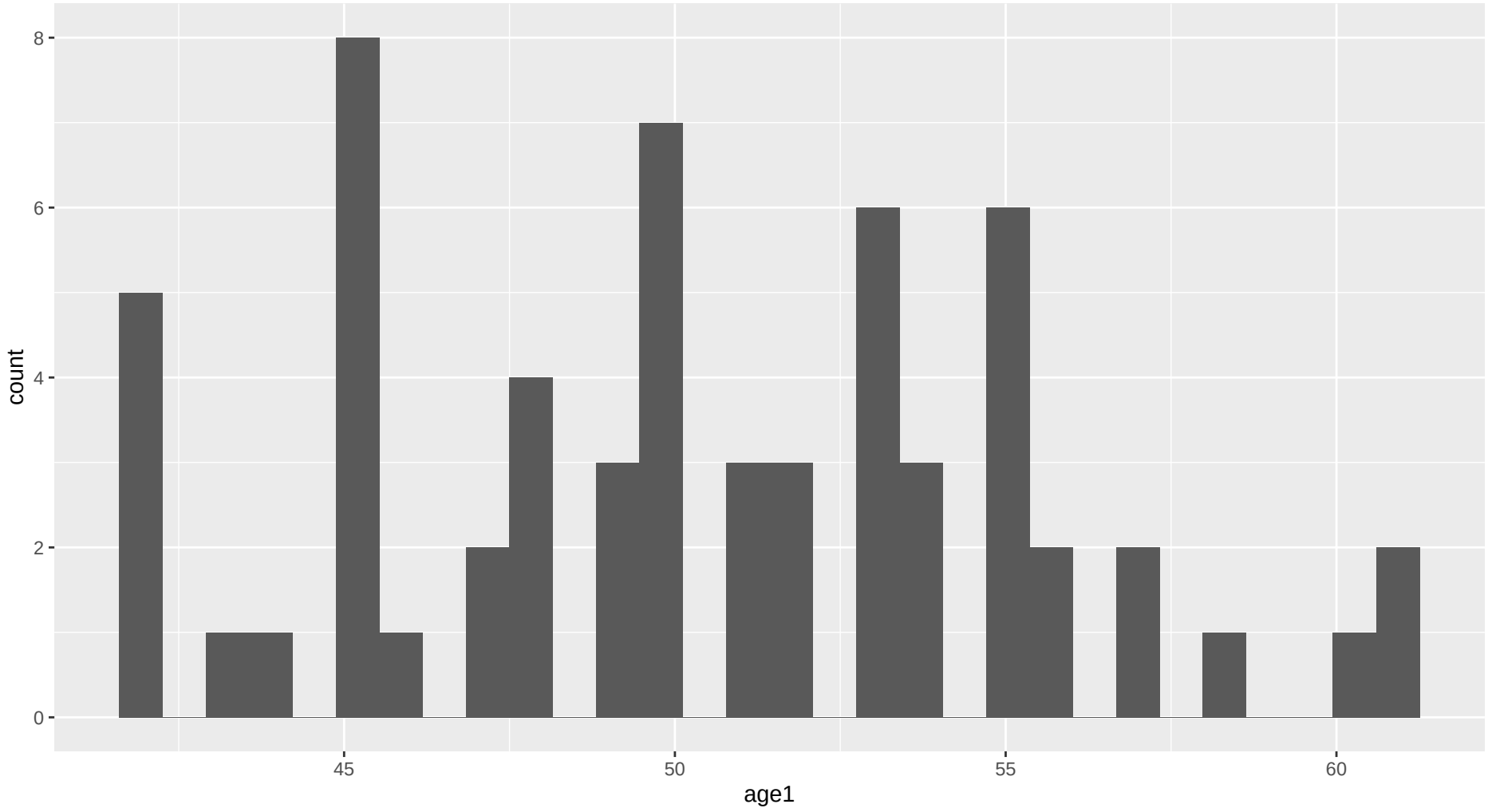
```
# A tibble: 61 x 3
   subject  age1  age2
   <chr>   <dbl> <dbl>
 1 S19-01     47    52
 2 S19-02     55    59
 3 S19-03     55    NA
 4 S19-04     45    45
 5 S19-05     45    48
 6 S19-06     42    49
 7 S19-07     43    55
 8 S19-08     50    46
 9 S19-09     54    50
10 S19-10     61    57
# ... with 51 more rows
```

Histogram of initial guesses?

```
ggplot(data = love_2019, aes(x = age1)) +
  geom_histogram()
```

Histogram of initial guesses?

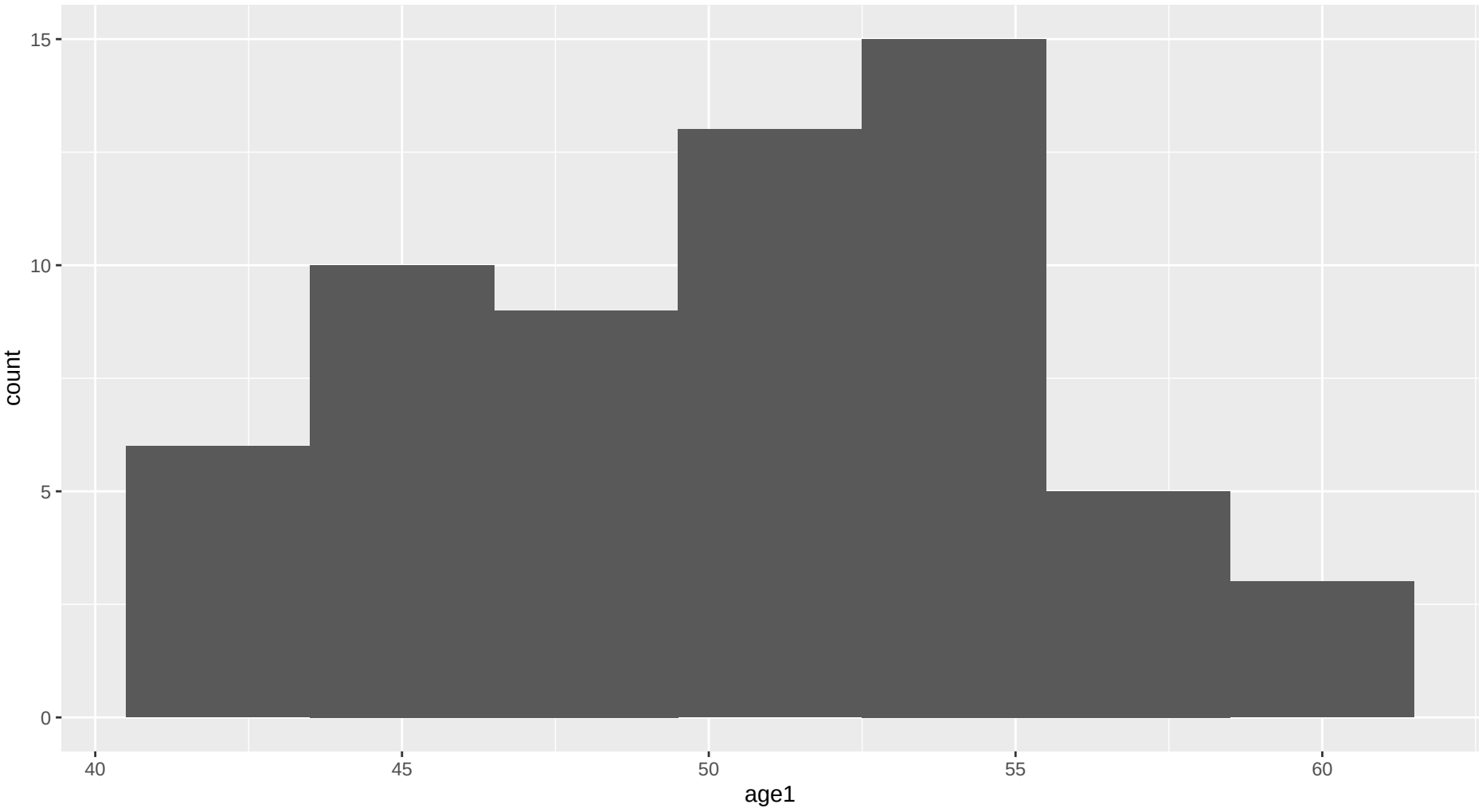
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.



Make the width of the bins 3 years?

```
ggplot(data = love_2019, aes(x = age1)) +
  geom_histogram(binwidth = 3)
```

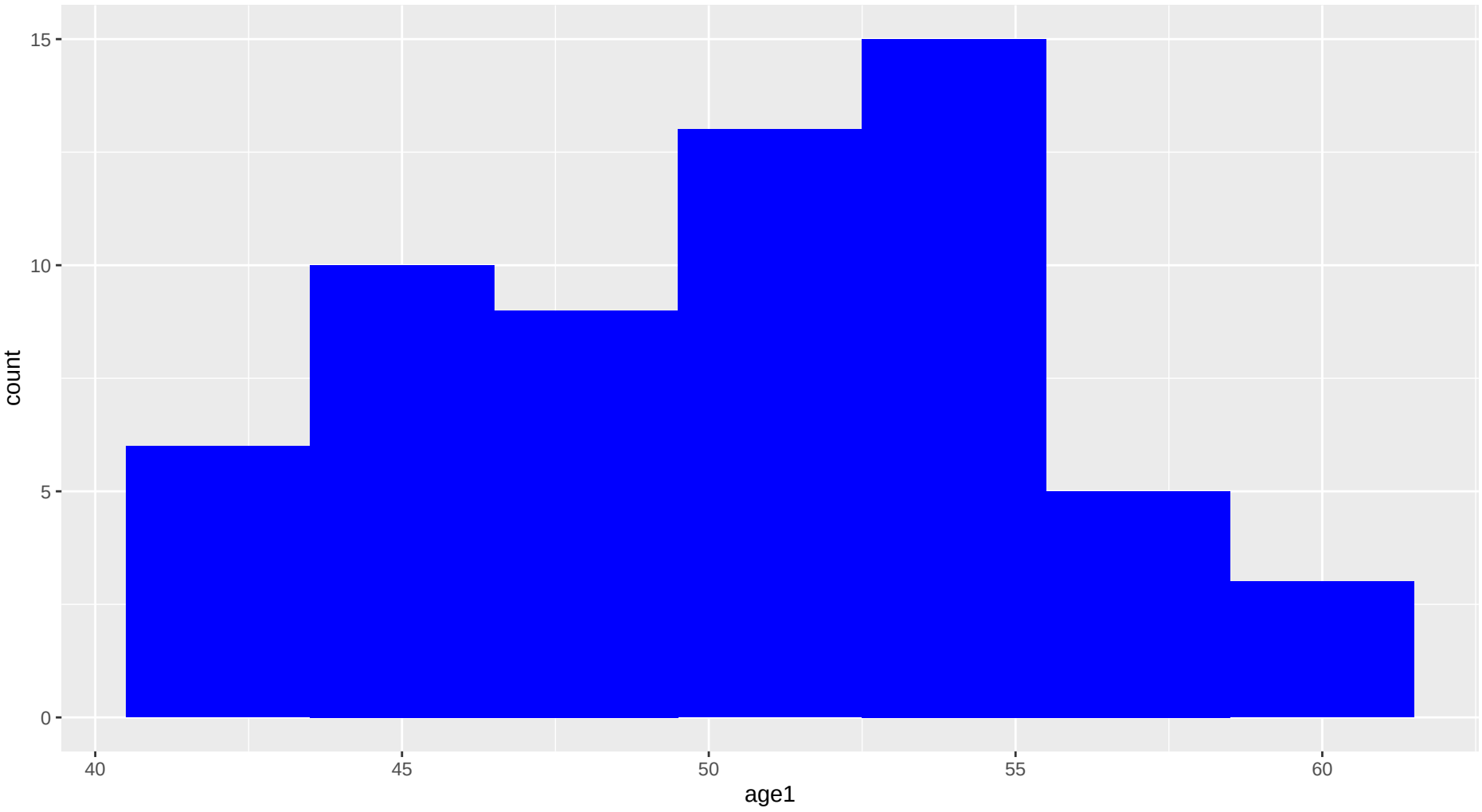
Make the width of the bins 3 years?



Fill in the bars with a better color?

```
ggplot(data = love_2019, aes(x = age1)) +
  geom_histogram(binwidth = 3,
                 fill = "blue")
```

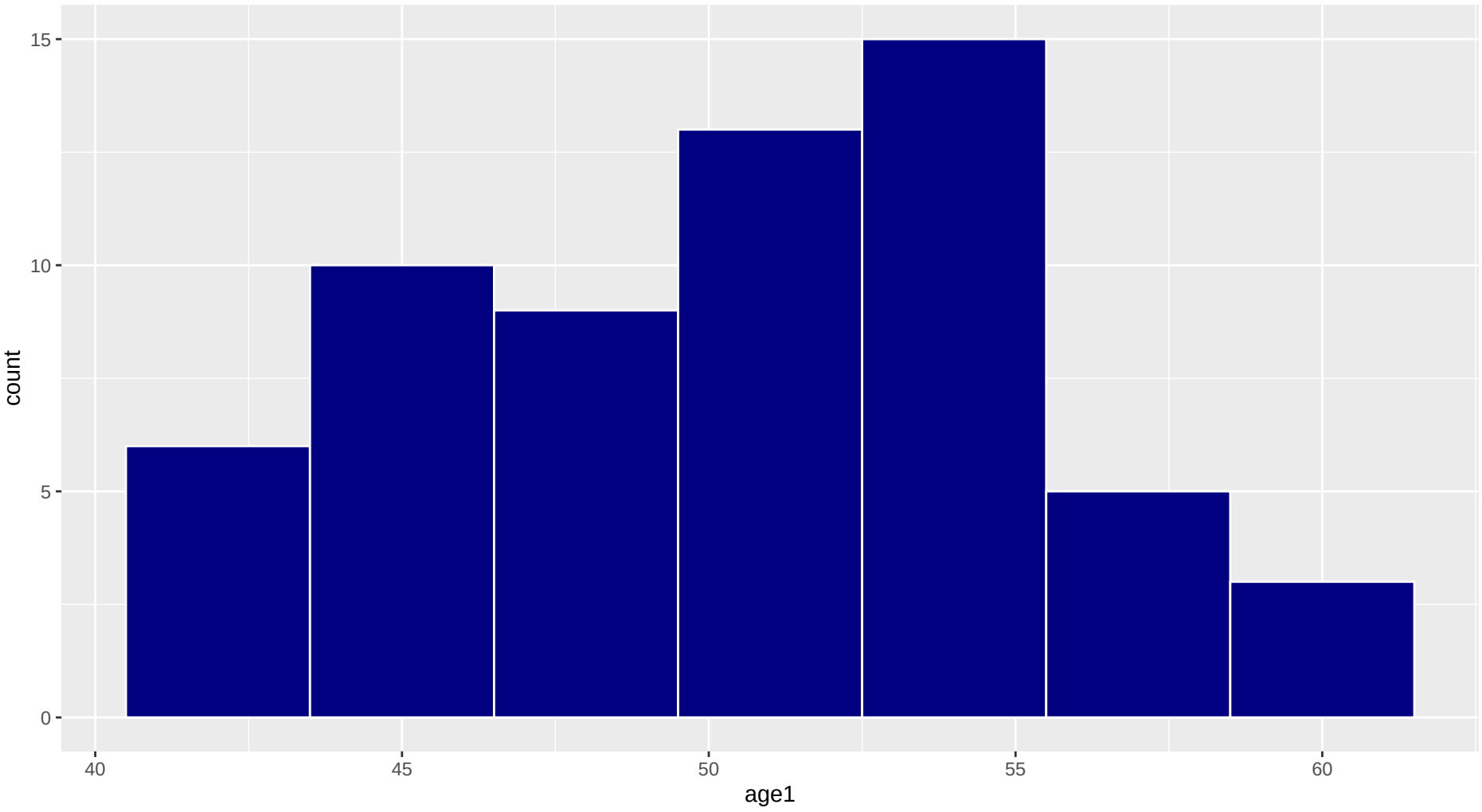
Fill in the bars with a better color?



Make it a little prettier?

```
ggplot(data = love_2019, aes(x = age1)) +
  geom_histogram(binwidth = 3,
                 fill = "navy", color = "white")
```

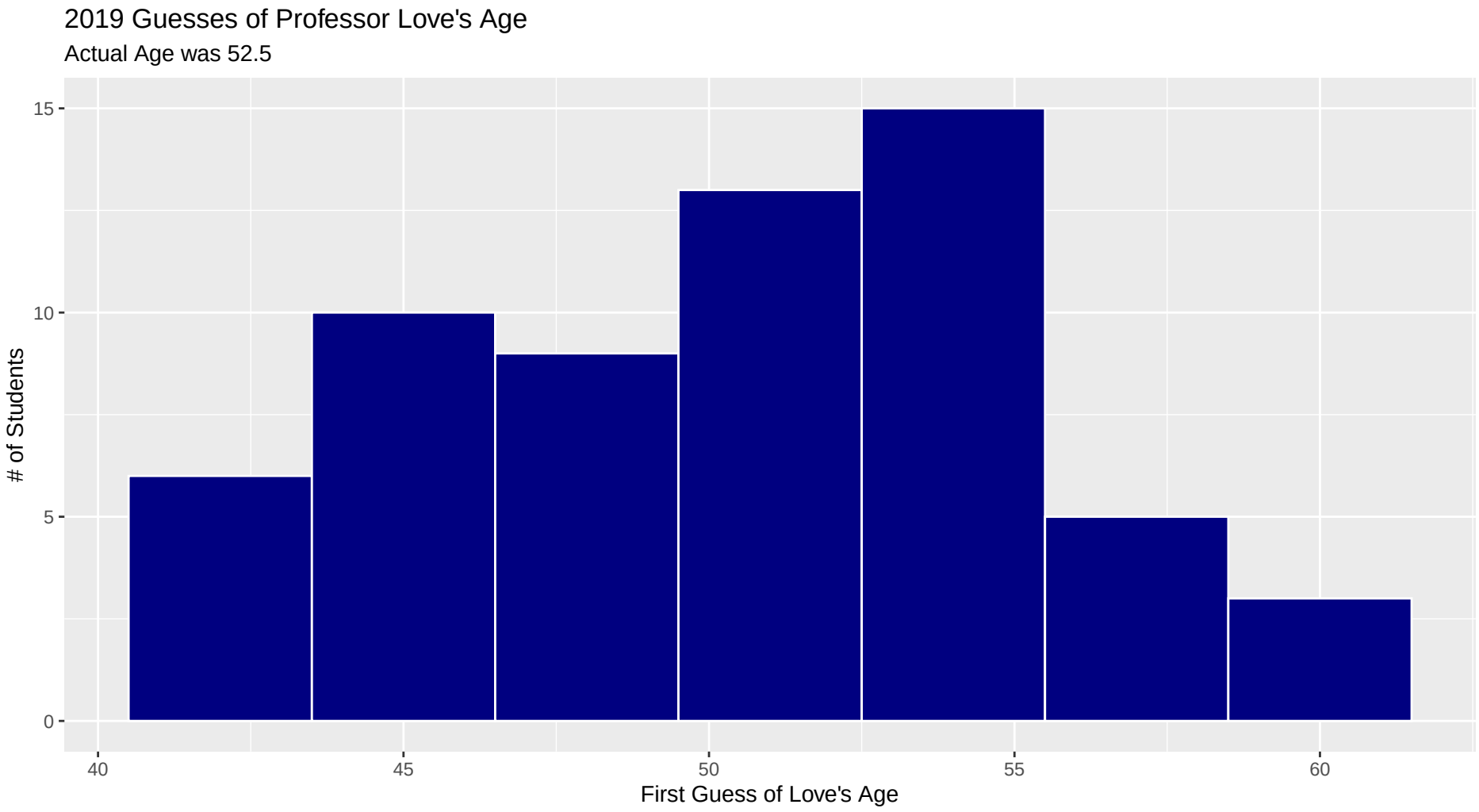
Make it a little prettier?



Add more meaningful labels?

```
ggplot(data = love_2019, aes(x = age1)) +
  geom_histogram(binwidth = 3,
                 fill = "navy", color = "white") +
  labs(x = "First Guess of Love's Age",
       y = "# of Students",
       title = "2019 Guesses of Professor Love's Age",
       subtitle = "Actual Age was 52.5")
```

Add more meaningful labels?



Numerical Summaries of Age Guesses

```r
summary(love_2019)
```

```
   subject                age1            age2      
 Length:61          Min.   :42.00   Min.   :42.00  
 Class :character   1st Qu.:46.00   1st Qu.:48.75  
 Mode  :character   Median :50.00   Median :52.00  
                    Mean   :50.34   Mean   :51.82  
                    3rd Qu.:54.00   3rd Qu.:55.00  
                    Max.   :61.00   Max.   :62.00  
                                    NA's   :1      
```

Some Additional Summaries

```
mosaic::favstats(~ age1, data = love_2019)
```

```
Registered S3 method overwritten by 'mosaic':
  method                           from
  fortify.SpatialPolygonsDataFrame ggplot2

 min Q1 median Q3 max     mean       sd  n missing
  42 46     50 54  61 50.34426 4.989607 61       0
```

```
mosaic::favstats(~ age2, data = love_2019)
```

```
 min    Q1 median Q3 max     mean       sd  n missing
  42 48.75     52 55  62 51.81667 4.545408 60       1
```

Another Approach

```
love_2019 %>%
  skimr::skim()
```

```
Skim summary statistics
 n obs: 61
 n variables: 3

-- Variable type:character ------------------------------------
 variable missing complete  n min max empty n_unique
  subject       0       61 61   6   6     0       61

-- Variable type:numeric --------------------------------------
 variable missing complete  n  mean   sd p0   p25 p50
     age1       0       61 61 50.34 4.99 42 46     50
     age2       1       60 61 51.82 4.55 42 48.75  52
 p75 p100     hist
  54   61 <U+2585><U+2586><U+2586><U+2587><U+2586><U+2587><U+2
```

A Better Look

```r
love_2019 %>%
  skimr::skim()
```

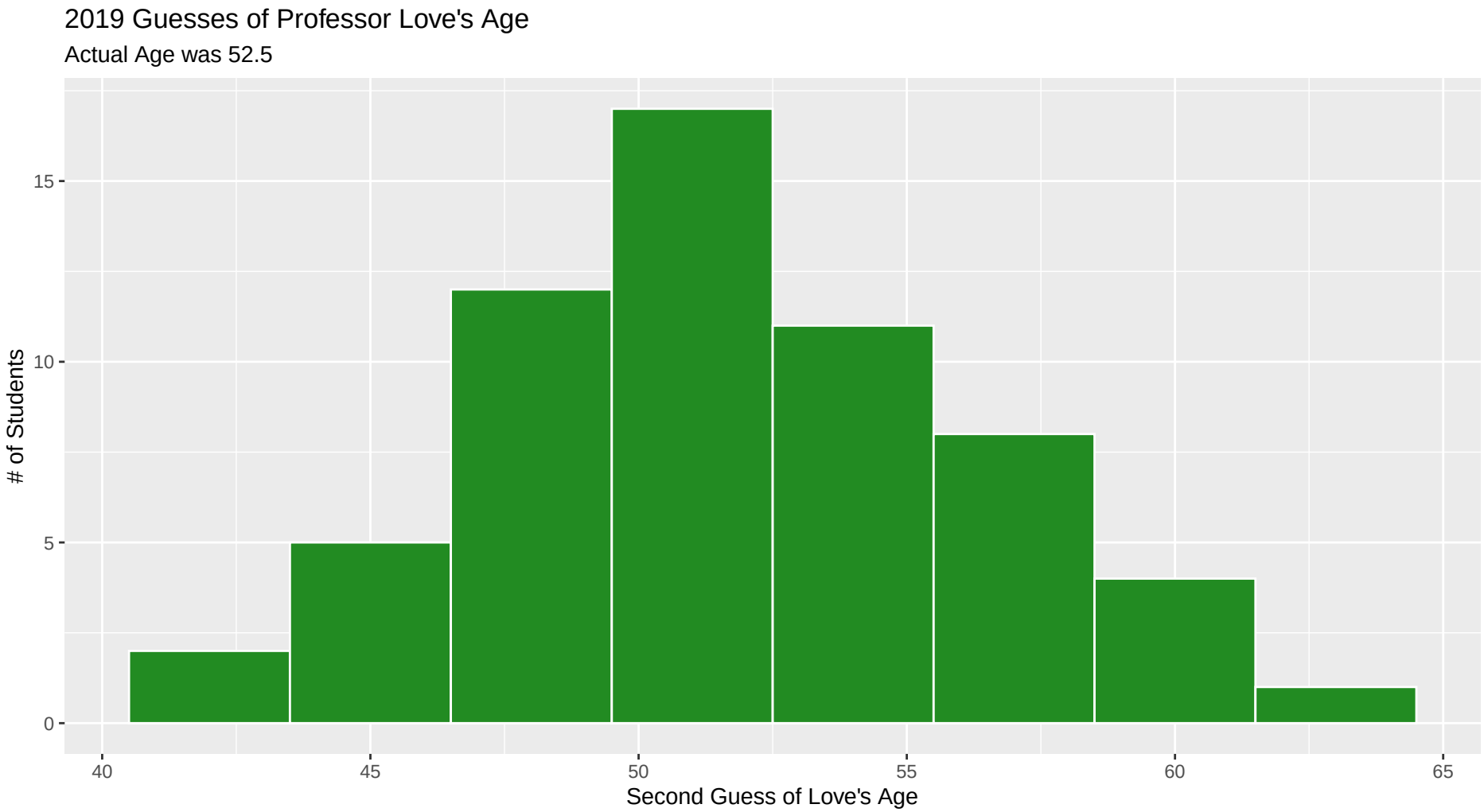
```
Skim summary statistics
 n obs: 61
 n variables: 3

-- Variable type:character -------------------------------------------------
 variable missing complete  n min max empty n_unique
  subject       0       61 61   6   6     0       61

-- Variable type:numeric ---------------------------------------------------
 variable missing complete  n  mean   sd p0   p25 p50 p75 p100     hist
     age1       0       61 61 50.34 4.99 42 46     50  54   61 ▂▃▅▇▅▅▃▂▁▁
     age2       1       60 61 51.82 4.55 42 48.75  52  55   62 ▂▃▅▇▅▅▃▂▁▁
```

What about the second guess?

```
ggplot(data = love_2019, aes(x = age2)) +
  geom_histogram(binwidth = 3,
                 fill = "forestgreen", color = "white") +
  labs(x = "Second Guess of Love's Age",
       y = "# of Students",
       title = "2019 Guesses of Professor Love's Age",
       subtitle = "Actual Age was 52.5")
```

What about the second guess?

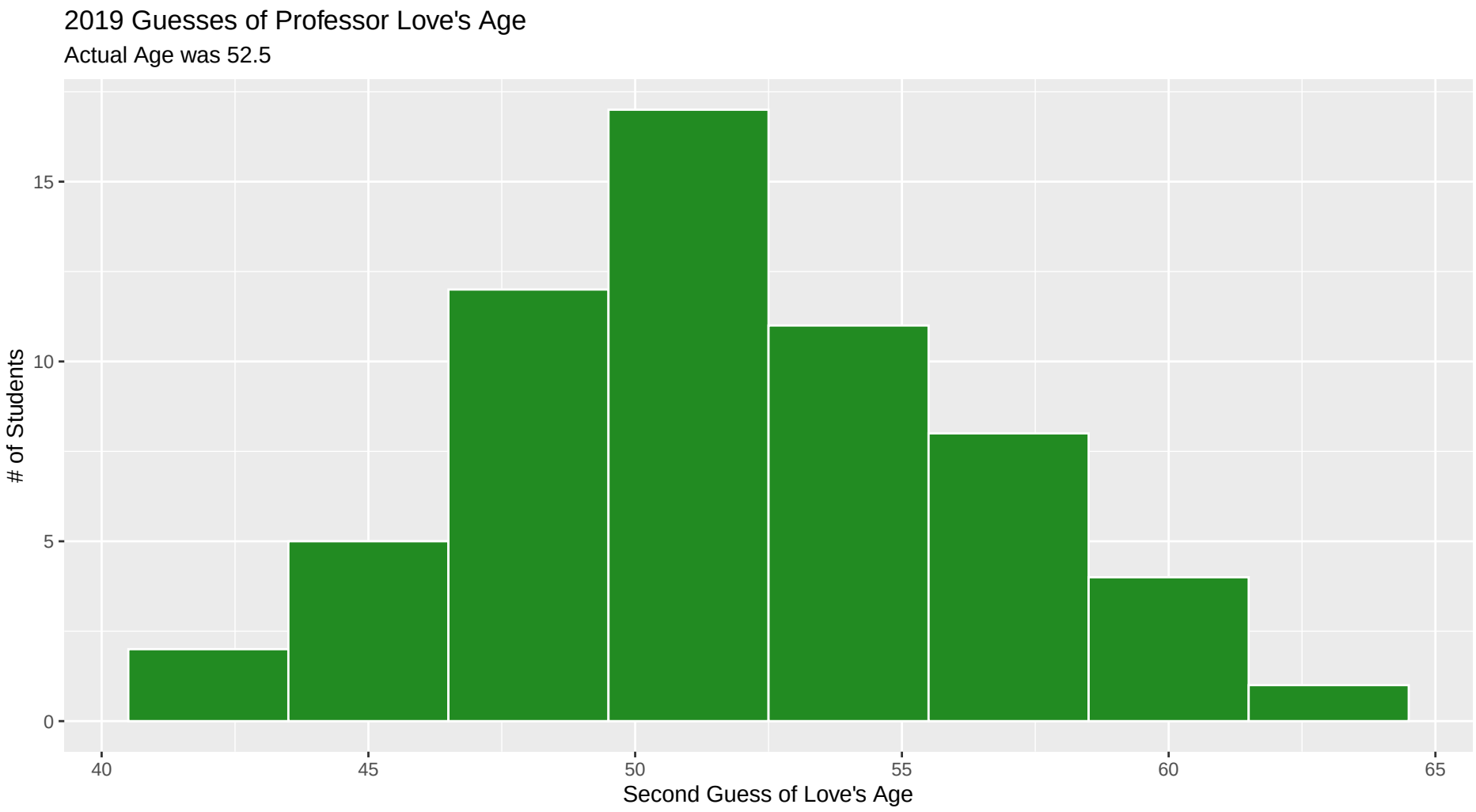
```
Warning: Removed 1 rows containing non-finite values
(stat_bin).
```



Filter to complete cases only

```
love_2019 %>%
  filter(complete.cases(age2)) %>%
  ggplot(data = ., aes(x = age2)) +
  geom_histogram(binwidth = 3,
                 fill = "forestgreen", color = "white") +
  labs(x = "Second Guess of Love's Age",
       y = "# of Students",
       title = "2019 Guesses of Professor Love's Age",
       subtitle = "Actual Age was 52.5")
```

Filter to complete cases only

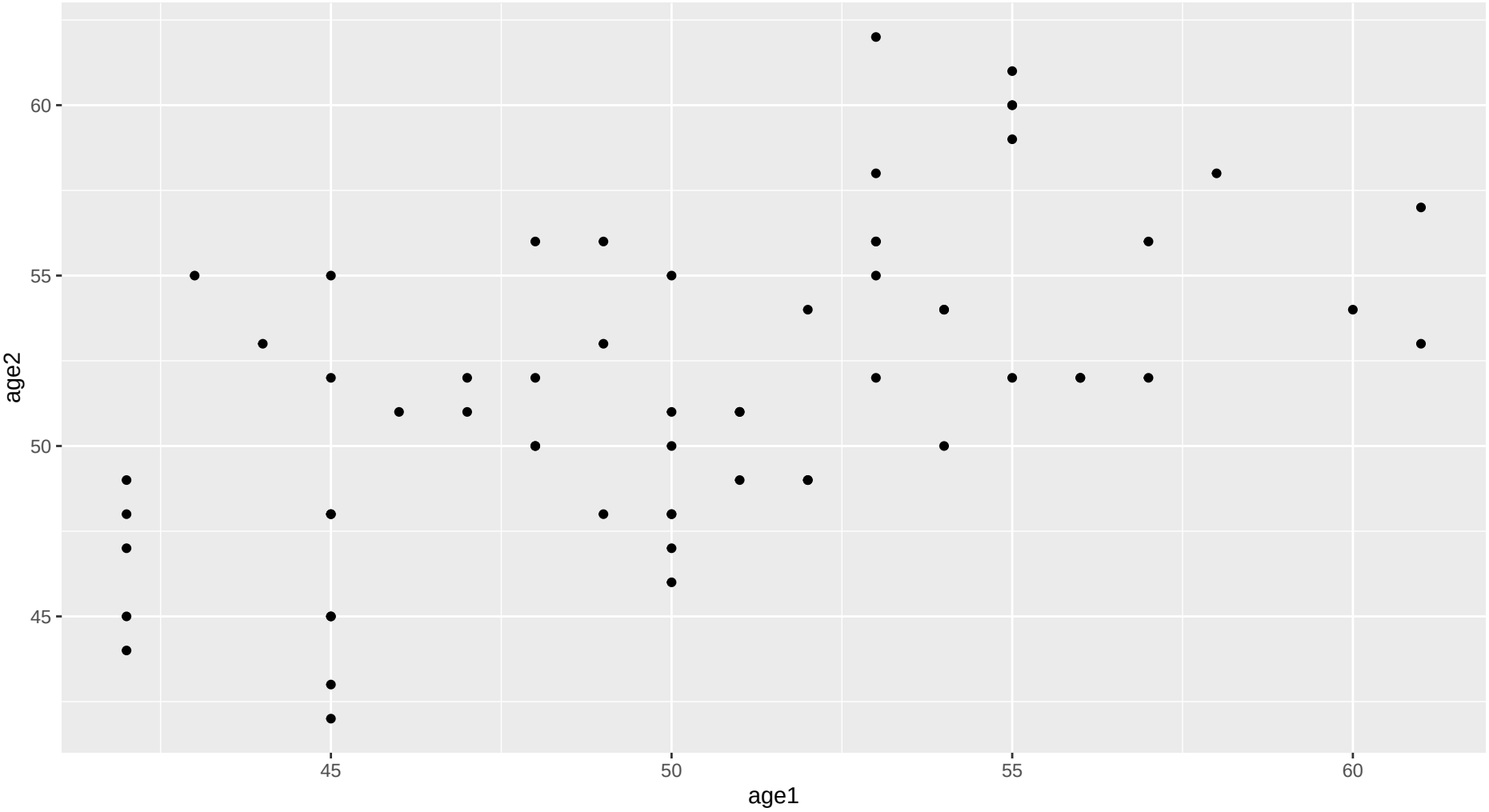


Comparing First Guess to Second Guess

```
ggplot(data = love_2019, aes(x = age1, y = age2)) +
  geom_point()
```

Comparing First Guess to Second Guess

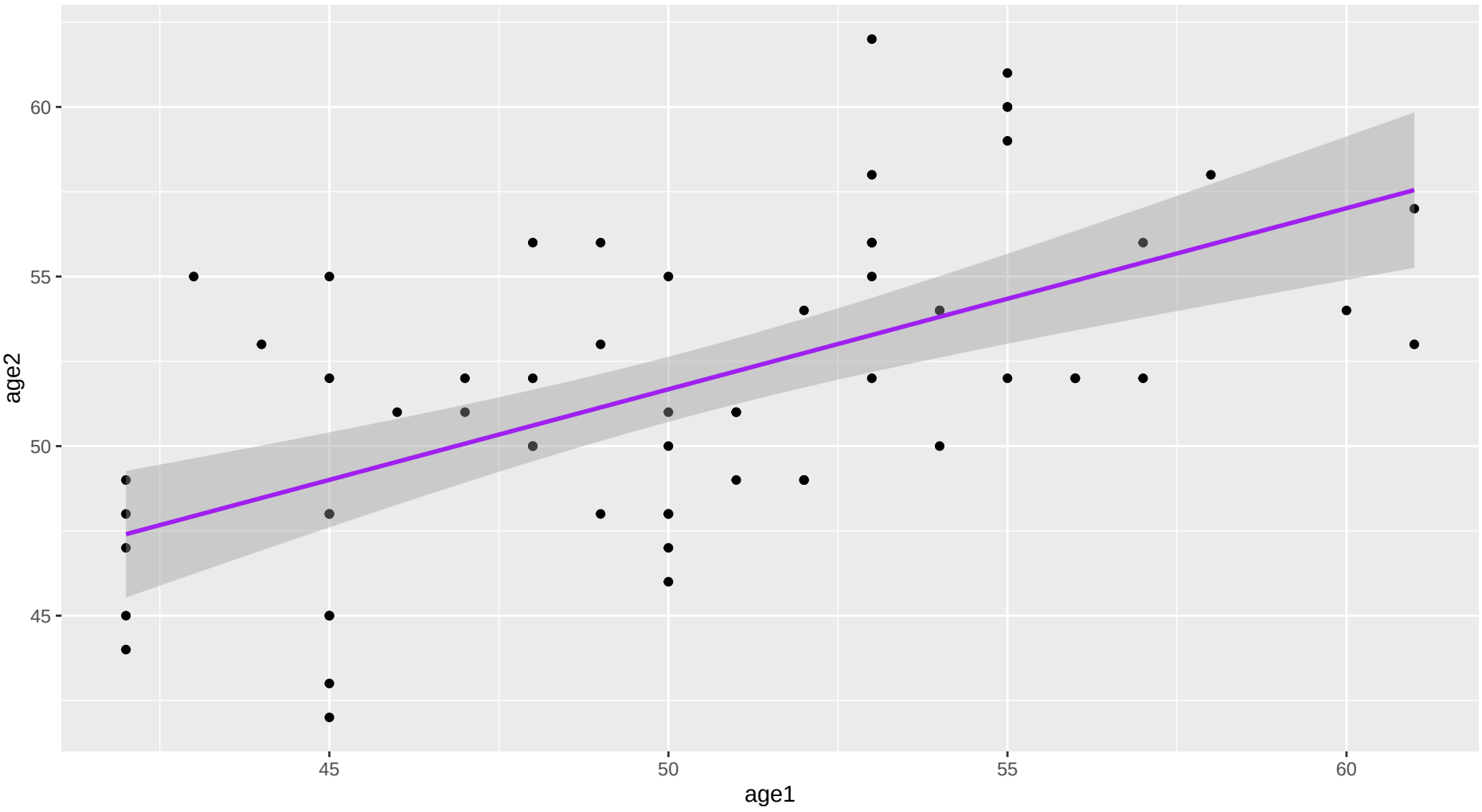
```
Warning: Removed 1 rows containing missing values
(geom_point).
```



Filter to complete cases, add regression line

```
love_2019 %>%
  filter(complete.cases(age1, age2)) %>%
  ggplot(data = ., aes(x = age1, y = age2)) +
  geom_point() +
  geom_smooth(method = "lm", col = "purple")
```

Filter to complete cases, add regression line



What's that regression line?

```r
lm(age2 ~ age1, data = love_2019)
```

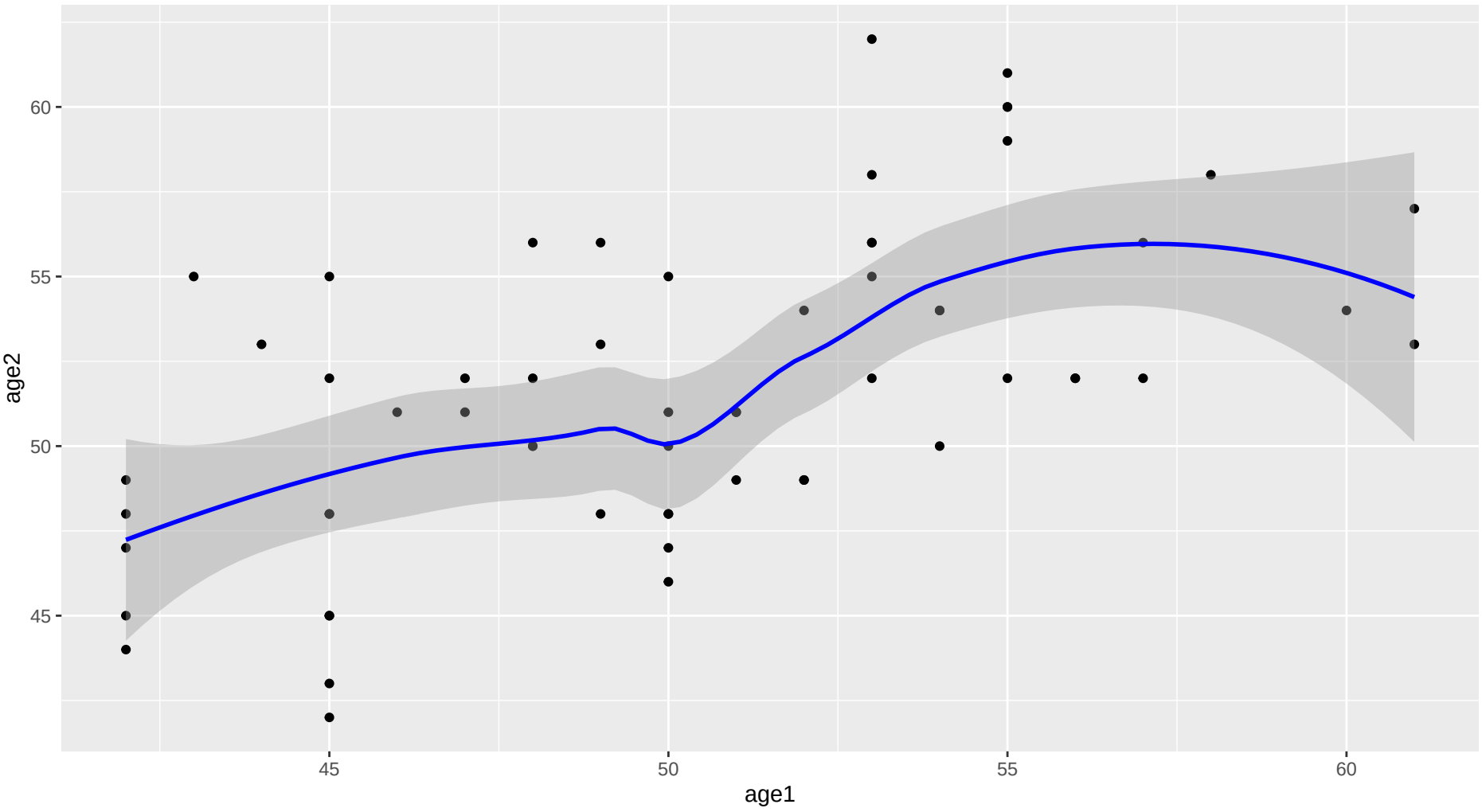
```
Call:
lm(formula = age2 ~ age1, data = love_2019)

Coefficients:
(Intercept)         age1
     24.973        0.534
```

- `lm` (by default) filters to complete cases.

How about a loess smooth curve, instead?

```
love_2019 %>%
  filter(complete.cases(age1, age2)) %>%
  ggplot(data = ., aes(x = age1, y = age2)) +
  geom_point() +
  geom_smooth(method = "loess", col = "blue")
```

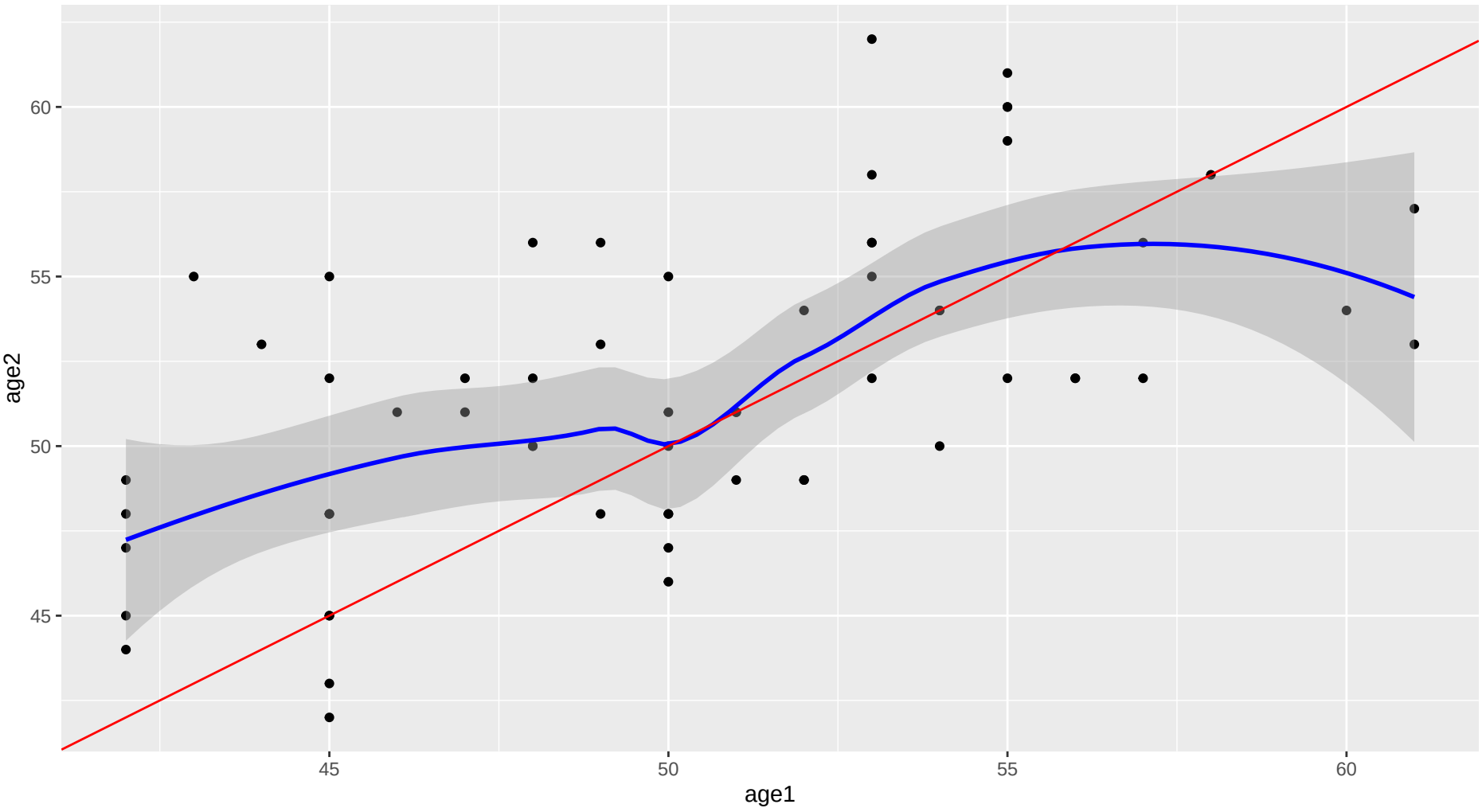
How about a loess smooth curve, instead?



Add a y = x line (no change in guess)?

```
love_2019 %>%
  filter(complete.cases(age1, age2)) %>%
  ggplot(data = ., aes(x = age1, y = age2)) +
  geom_point() +
  geom_smooth(method = "loess", col = "blue") +
  geom_abline(intercept = 0, slope = 1, col = "red")
```

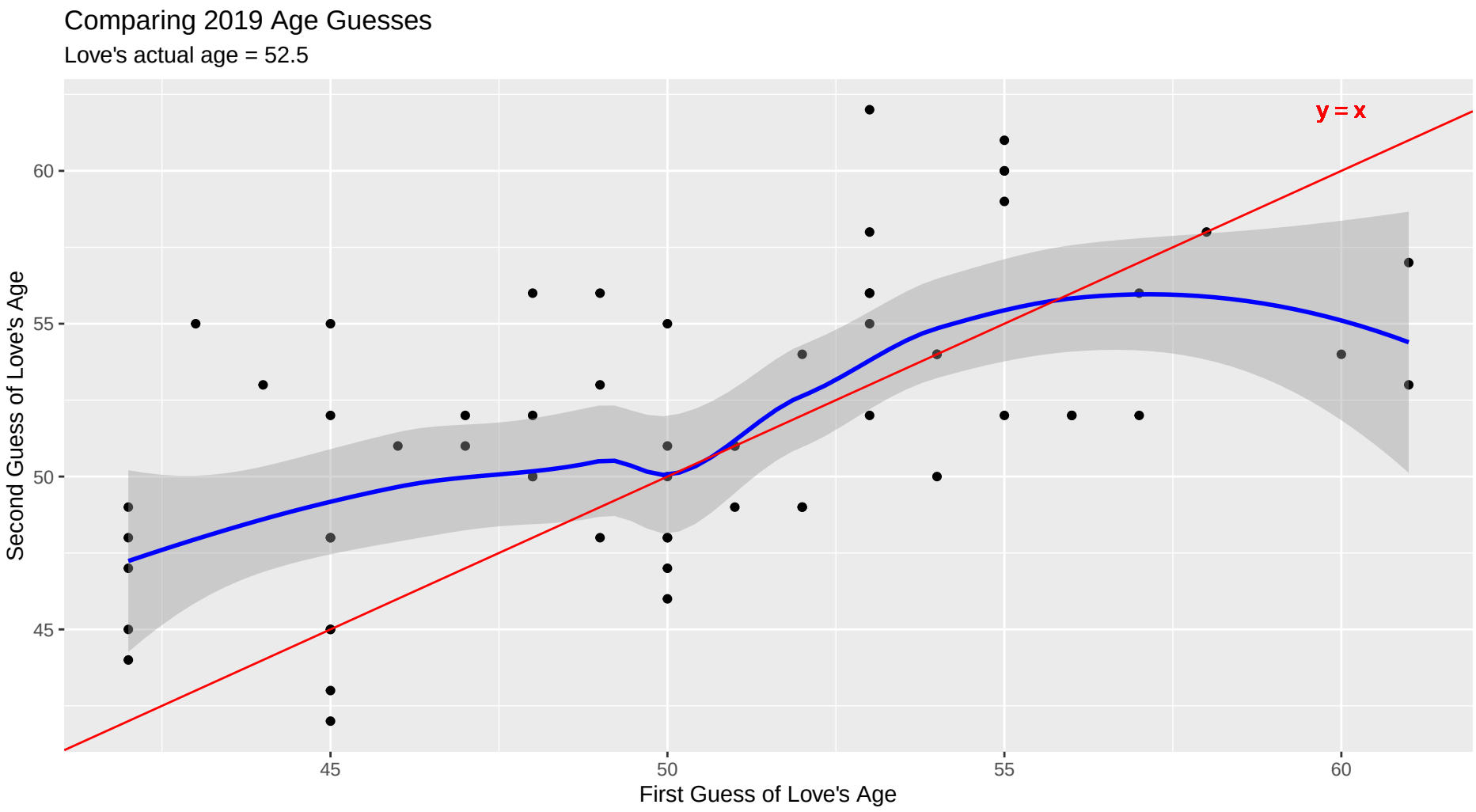
Add a y = x line (no change in guess)?



Add more meaningful labels

```
love_2019 %>%
  filter(complete.cases(age1, age2)) %>%
  ggplot(data = ., aes(x = age1, y = age2)) +
  geom_point() +
  geom_smooth(method = "loess", col = "blue") +
  geom_abline(intercept = 0, slope = 1, col = "red") +
  geom_text(x = 60, y = 62,
            label = "y = x", col = "red") +
  labs(x = "First Guess of Love's Age",
       y = "Second Guess of Love's Age",
       title = "Comparing 2019 Age Guesses",
       subtitle = "Love's actual age = 52.5")
```

Add more meaningful labels



age1 - age2 difference in guesses?

```
love_2019 %>%
  mutate(diff = age1 - age2) %>%
  skimr::skim()
```

```
Skim summary statistics
 n obs: 61
 n variables: 4

-- Variable type:character -------------------------------------------------
 variable missing complete  n min max empty n_unique
  subject       0       61 61   6   6     0       61

-- Variable type:numeric ---------------------------------------------------
 variable missing complete  n  mean   sd  p0   p25 p50 p75 p100     hist
     age1       0       61 61 50.34 4.99  42 46     50  54   61
     age2       1       60 61 51.82 4.55  42 48.75  52  55   62
     diff       1       60 61 -1.55 4.35 -12 -5     -2   2    8
```

How Many Guesses Increased?

```
love_2019 %>%
  mutate(diff = age1 - age2) %>%
  count(diff < 0)
```

```
# A tibble: 3 x 2
  `diff < 0`     n
  <lgl>      <int>
1 FALSE         28
2 TRUE          32
3 NA             1
```

Increased / Stayed the Same / Decreased

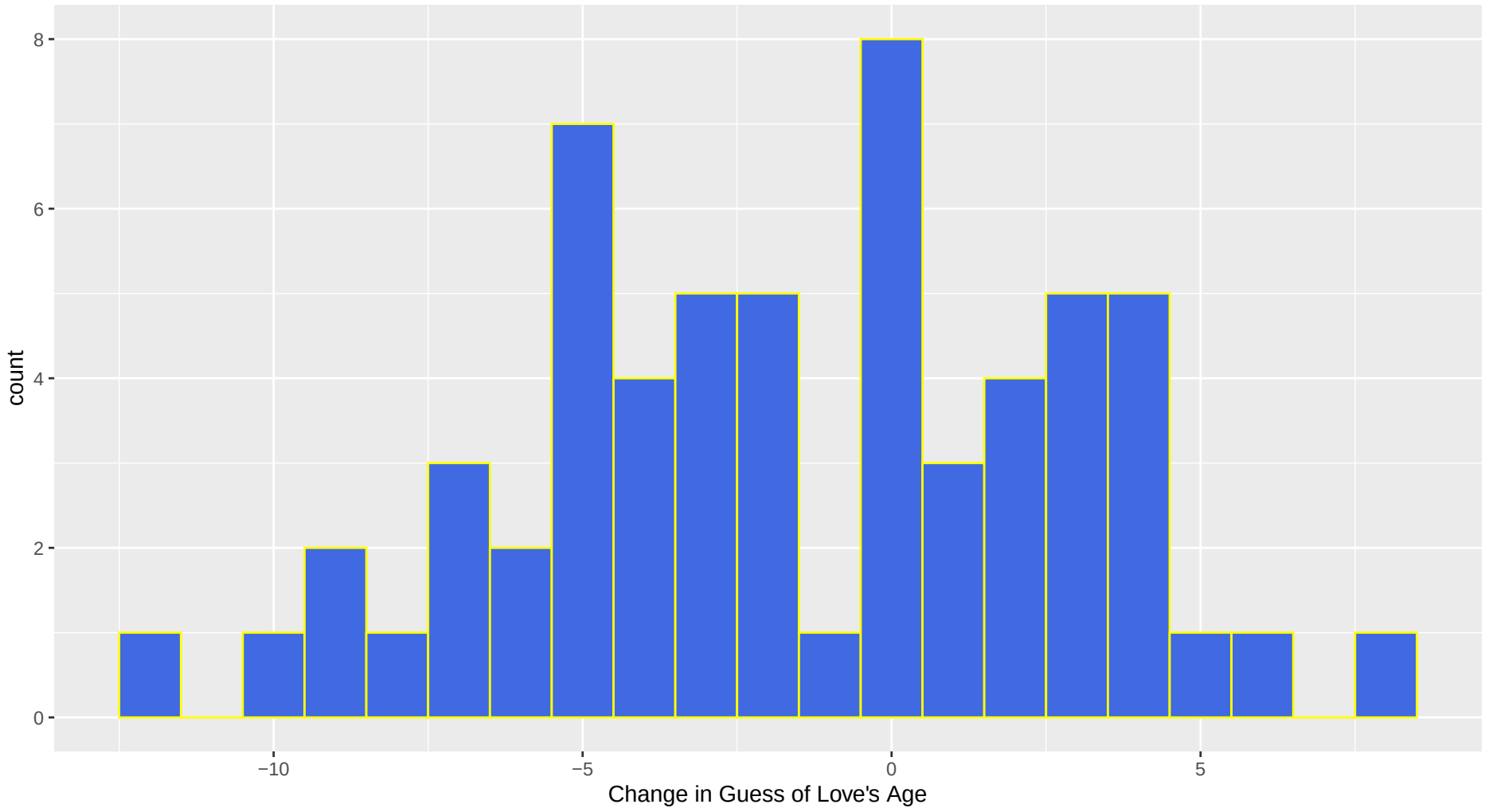
```
love_2019 %>%
  mutate(diff = age1 - age2) %>%
  count(sign(diff))
```

```
# A tibble: 4 x 2
  `sign(diff)`     n
         <dbl> <int>
1           -1    32
2            0     8
3            1    20
4           NA     1
```

Histogram of difference in guesses

```
love_2019 %>%
  mutate(diff = age1 - age2) %>%
  filter(complete.cases(diff)) %>%
  ggplot(data = ., aes(x = diff)) +
  geom_histogram(binwidth = 1,
                 fill = "royalblue", color = "yellow") +
  labs(x = "Change in Guess of Love's Age")
```

Histogram of difference in guesses



Analyzing the Survey Data - A little challenge

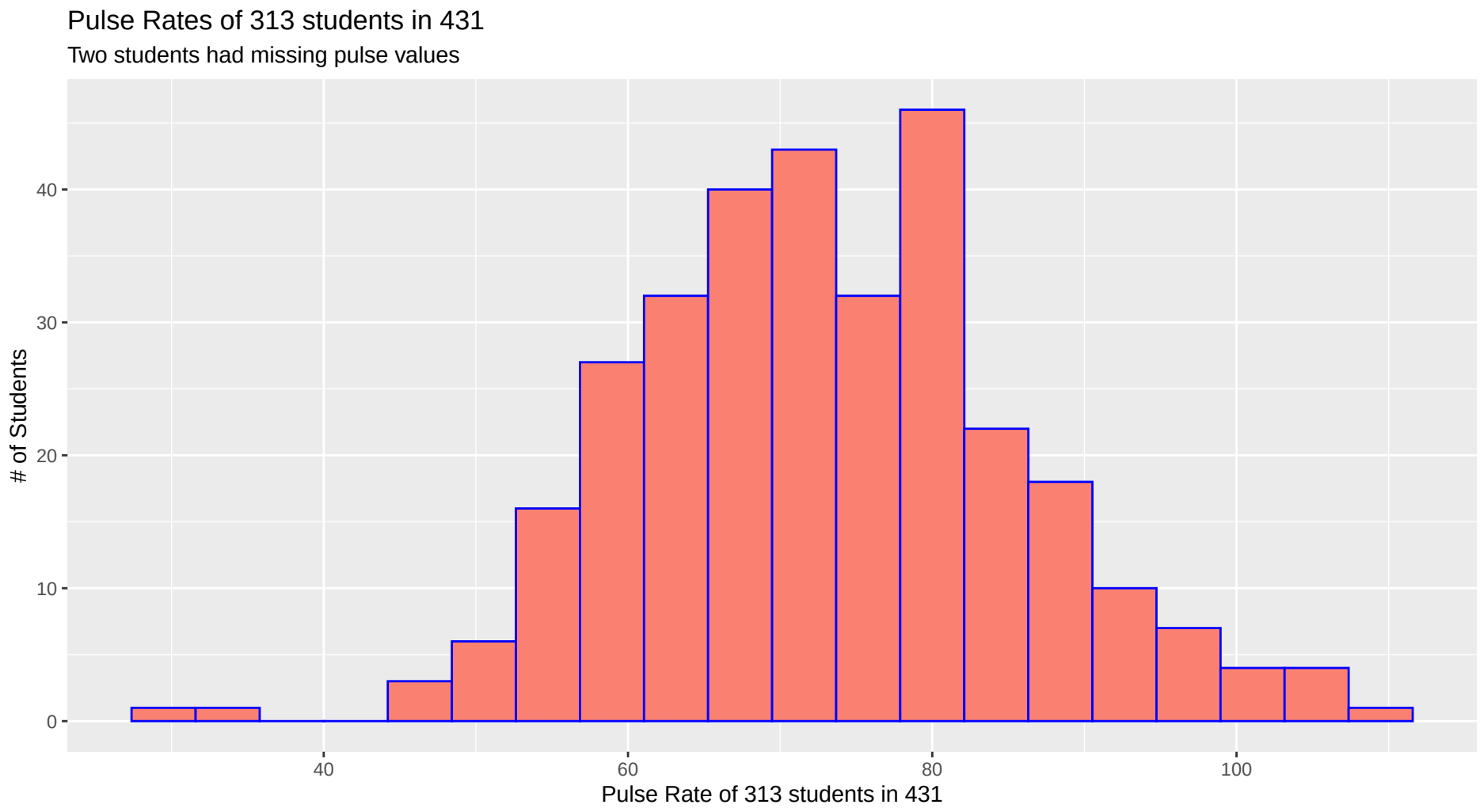
We have data on the site in a file called `surveyday1_2019.csv`. Build a project to study those data.

Put the data in a file called `survey1` in R.

- I'd call my R Markdown file `day1surveyanalysis`

Can you reproduce the following...

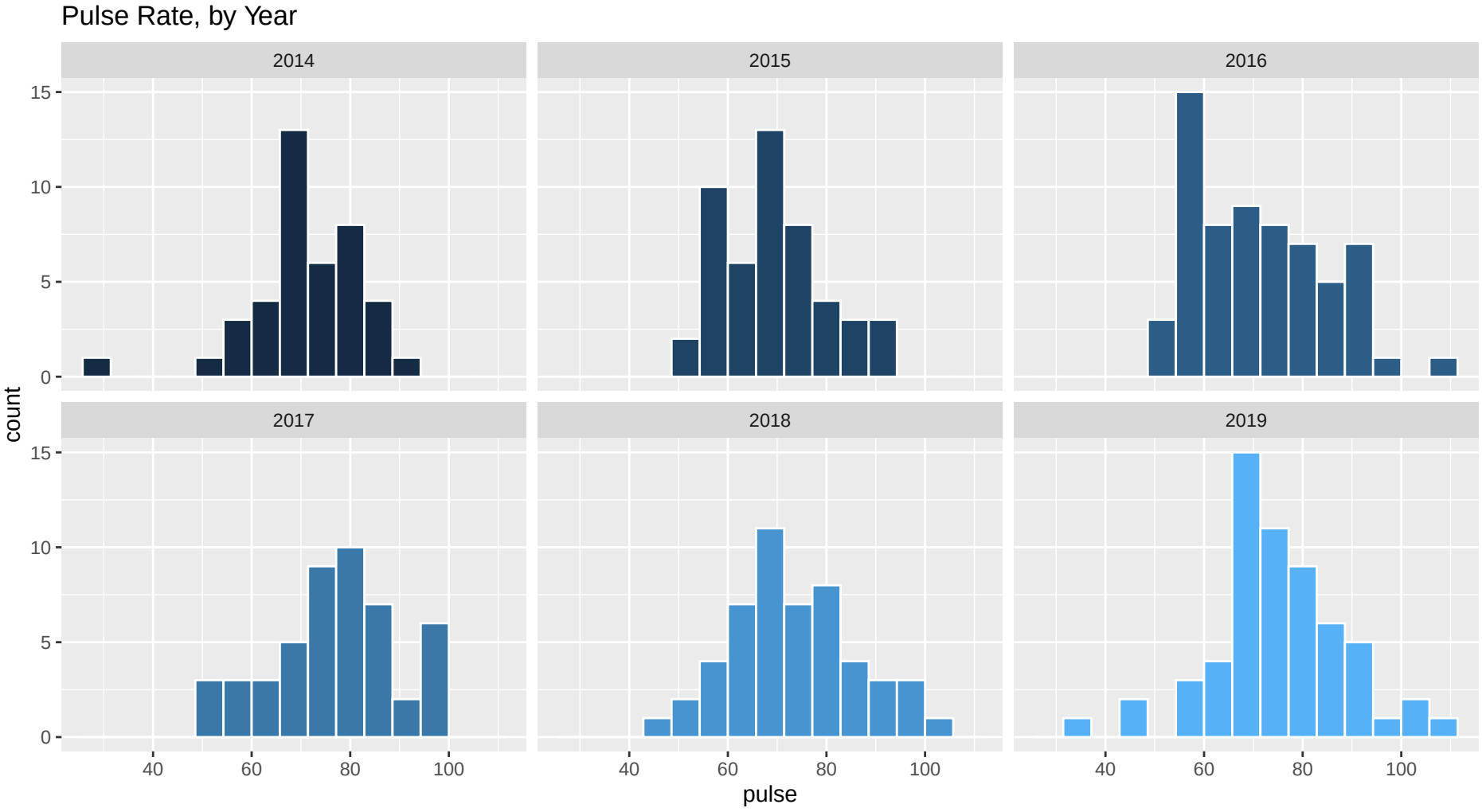
A. That fill color is called *salmon*, I used 20 bins.



Code for Part A.

```
survey1 <- read_csv("data/surveyday1_2019.csv")

survey1 %>% filter(complete.cases(pulse)) %>%
  ggplot(data = ., aes(x = pulse)) +
  geom_histogram(bins = 20, col = "blue", fill = "salmon") +
  labs(x = "Pulse Rate of 313 students in 431",
       y = "# of Students",
       title = "Pulse Rates of 313 students in 431",
       subtitle = "Two students had missing pulse values")
```

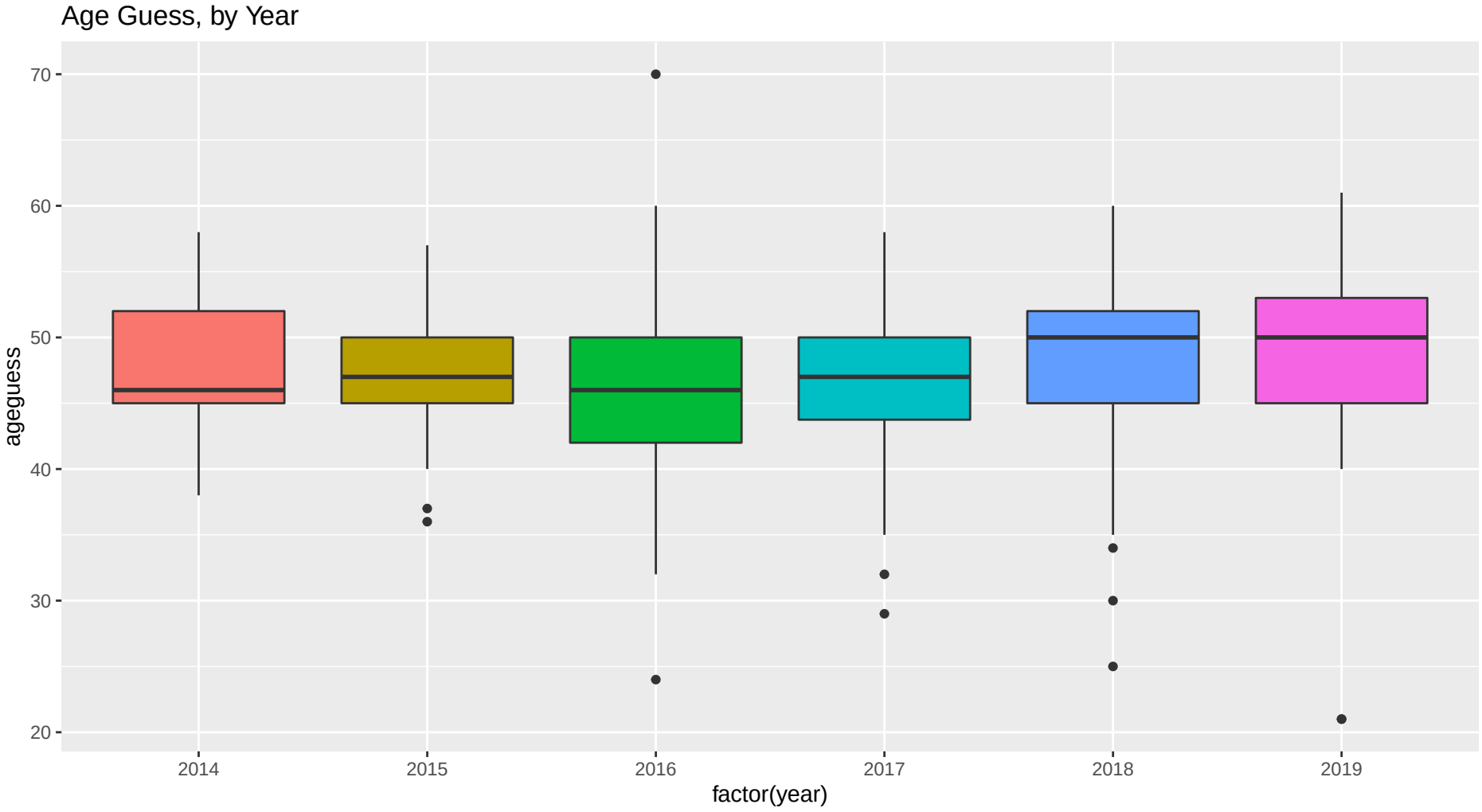
B. Histograms of Pulse Rates, Faceted by Year



Code for Plot B.

```
survey1 %>% filter(complete.cases(pulse)) %>%
  ggplot(data = ., aes(x = pulse, fill = year)) +
  geom_histogram(bins = 15, col = "white") +
  facet_wrap(~ year) +
  guides(fill = FALSE) +
  labs(title = "Pulse Rate, by Year")
```

C. Boxplots of Age Guesses, by Year



Code for Plot C

```
survey1 %>% filter(complete.cases(ageguess)) %>%
  ggplot(data = ., aes(x = factor(year), y = ageguess,
                       fill = factor(year))) +
  geom_boxplot() +
  guides(fill = FALSE) +
  labs(title = "Age Guess, by Year")
```

Summary Table of Age Guesses, by Year

```
# A tibble: 6 x 5
   year     n  mean    sd median
  <dbl> <int> <dbl> <dbl>  <dbl>
1  2014    41  47.3  5.21     46
2  2015    49  47.1  4.62     47
3  2016    61  46.0  7.00     46
4  2017    48  46.5  6.15     47
5  2018    50  48.2  6.47     50
6  2019    60  48.6  7.09     50
```

Code for Summary Table

```
survey1 %>%
  filter(complete.cases(ageguess)) %>%
  group_by(year) %>%
  summarize(n = n(),
            mean = mean(ageguess, na.rm=TRUE),
            sd = sd(ageguess, na.rm=TRUE),
            median = median(ageguess, na.rm=TRUE)
            )
```