431 Quiz 1 for Fall 2018

Thomas E. Love

due 2018-10-08 at 7 PM, version 2018-10-09

Contents

General Instructions

Please select or type in your best response (or responses, as indicated) for each question. The deadline for completing the quiz is 7 PM on Monday 2018-10-08, and this is a firm deadline.

The questions are not arranged in any particular order, and your score is based on the number of correct responses, so you should answer all questions. There are 40 questions, and each question is worth 2, 2.5 or 3 points. The maximum possible score on the quiz is 100 points.

If you wish to work on some of the quiz and then return to work on the rest of the quiz or edit your responses later, you can do this by [1] completing the final question which asks you to type in your full name, and then [2] submit the quiz. You will then receive a link which allows you to return to the quiz without losing your progress.

Occasionally, I ask you to provide a single line of code. In all cases, a single line of code can include at most one pipe for these purposes, although you may or may not need the pipe in any particular setting. Moreover, you need not include the library command at any time for any of your code. We will assume the relevant packages have been loaded in R.

You are welcome to consult the materials provided on the course website, but you are not allowed to discuss the questions on this quiz with anyone other than Professor Love and the teaching assistants at 431-help at case dot edu. Please submit any questions you have about the Quiz to 431-help through email. Thank you, and good luck.

Question 01

Suppose that the height (in cm) of adult non-hispanic white women living in the state of Ohio follows a Normal model with mean 163 and standard deviation 7. If this is the case, then what percentage of adult non-hispanic white women living in the state of Ohio would have a height of 156 cm or larger? Please round your response to the nearest integer.

Answer 01 is 84

I'd hoped you'd simply realize that 16% of data falls more than one standard deviation below the mean, and thus save yourself a calculation, since the remaining 84% must fall above that level, but if not, you certainly could use `pnorm` in R to do the calculation.

```
pnorm(156, mean = 163, sd = 7, lower.tail = FALSE)
```

```
[1] 0.8413447
```

Results on Q01, worth 2 points.

- There are 51 students who took the Quiz.
- At least 46 gave the correct response, so at least 90%. Sensational!
- A common incorrect response was 16, and I can guess where that comes from.
- There was no partial credit for this question.

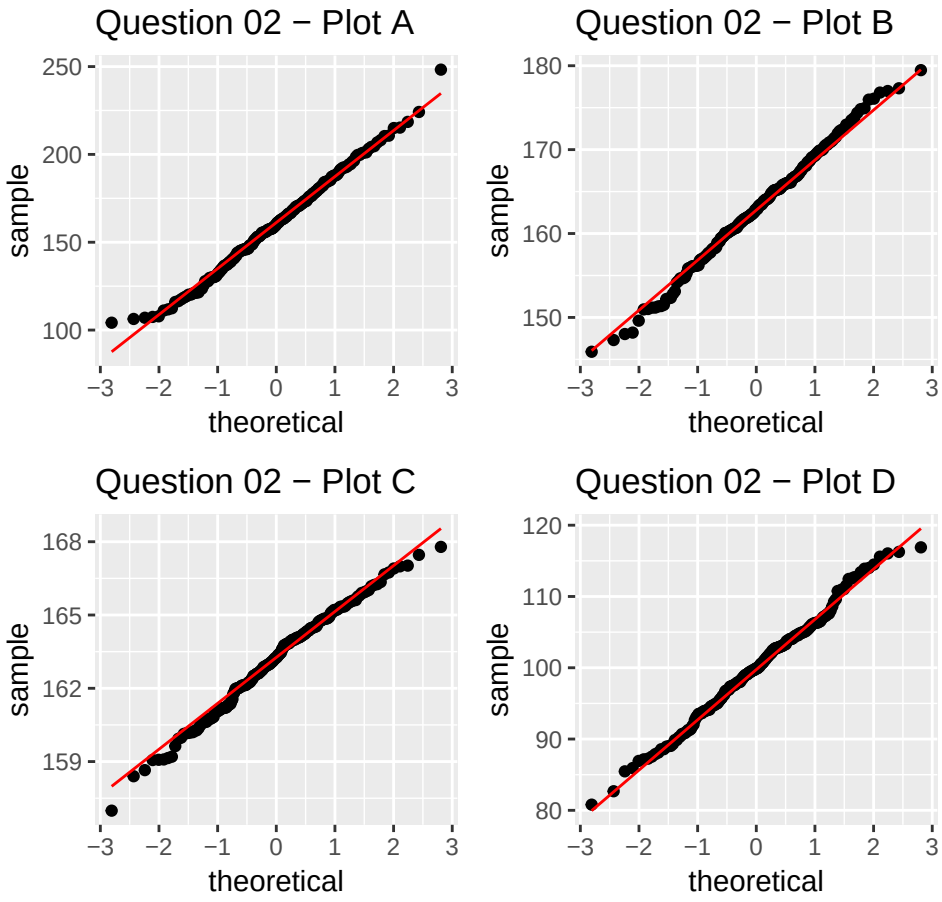
Question 02

There are four plots shown in the Figure for Question 02. Each shows a Normal Q-Q plot describing a different set of 200 heights. Which of the plots below shows data that could plausibly come from the Normal model specified in Question 01?

The response options for Question 02 are:

a. Plot A
b. Plot B
c. Plot C
d. Plot D
e. None of the above.

Figure for Question 02



Answer for Question 02 is b

The Normal model of interest has mean 163 and standard deviation 7. Each of the four plots is reasonably close to a Normal distribution in terms of shape. So, let's think about the implications of a Normal model with mean 163 and standard deviation 7. For example, almost all of the data should fall within 3 standard deviations of the mean (from 163-21 = 139 to 163+21 = 184, roughly).

Looking at the Y axis in each plot shows us the range of each sample. Plot A displays a much larger standard deviation than 7, and plot C displays a much smaller standard deviation than 7. Plot D appears to have a standard deviation close to the model's 7, but the mean is much smaller than 163. Only Plot B has an

appropriate mean near 163 and standard deviation near 7, and in fact Plot B is the only data set derived from the Normal model specified in Question 01.

Results on Q02, worth 3 points.

- 38/51 (75%) gave the correct response.
- There was no partial credit for this question.

Response	a	b	c	d
Count	3	**38**	4	5

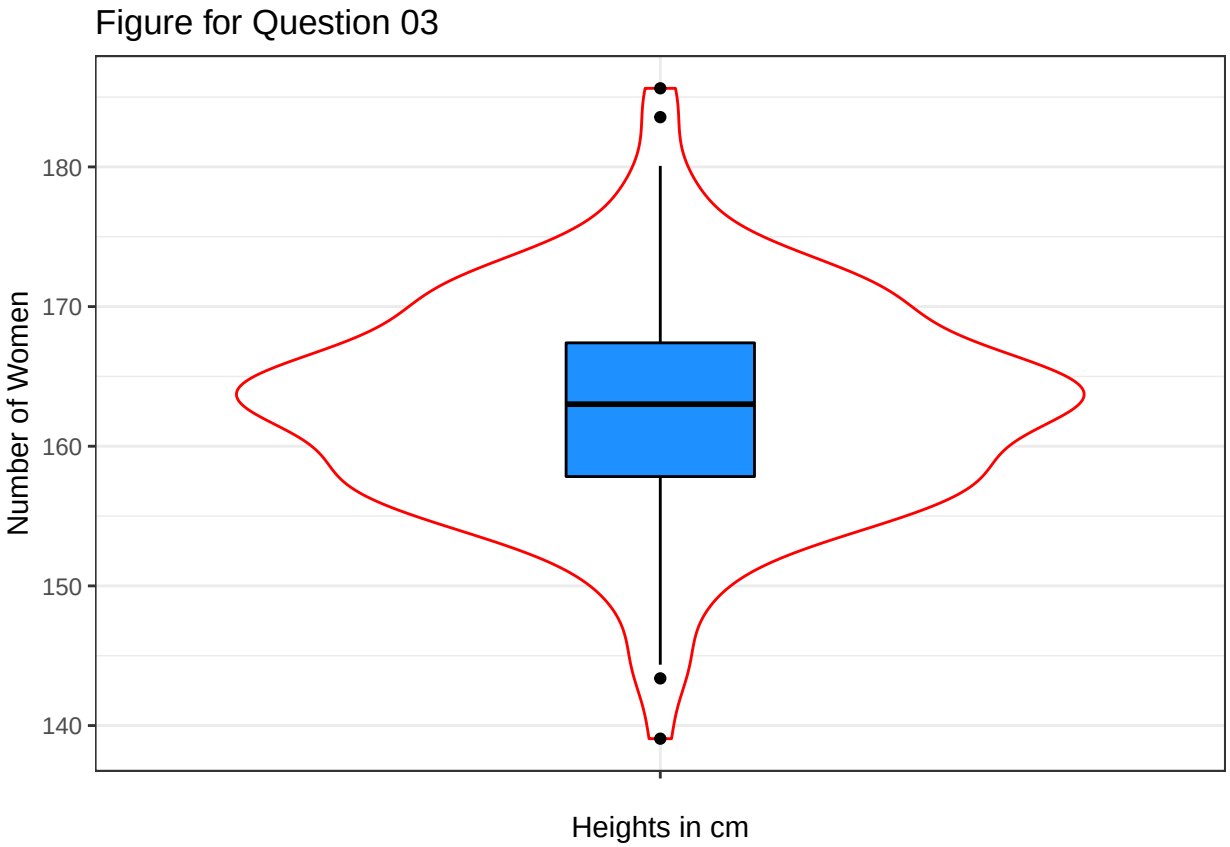
Question 03

A new sample of heights for 148 adult non-hispanic white women living in the state of Ohio have been stored in the `height` column in a tibble called `dat03`. Which of the following bits of R code was NOT used in generating the Figure for Question 03?

The response options for Question 03 are:

a. labs(title = "Figure for Question 03")
b. ggplot(data = dat03, aes(x = "", y = height))
c. geom_boxplot(fill = "dodgerblue", col = "black", width = 0.2)
d. geom_bar(fill = "blue", col = "white")
e. labs(x = "Heights in cm", y = "Number of Women")
f. geom_violin(col = "red")
g. theme_bw()

Figure for Question 03



Answer for Question 03 is d

The Figure in Question 03 was made by combining the other six bits of code. This is a boxplot and a violin plot, but there's no bar chart here.

Here's the actual code that was used. . .

```r
set.seed(2018003)
temp <- rnorm(148, mean = 163.3, sd = 8)

dat03 <- data_frame(height = temp)

ggplot(data = dat03, aes(x = "", y = height)) +
    geom_violin(col = "red") +
    geom_boxplot(fill = "dodgerblue", col = "black", width = 0.2) +
    labs(x = "Heights in cm", y = "Number of Women") +
    labs(title = "Figure for Question 03") +
    theme_bw()
```

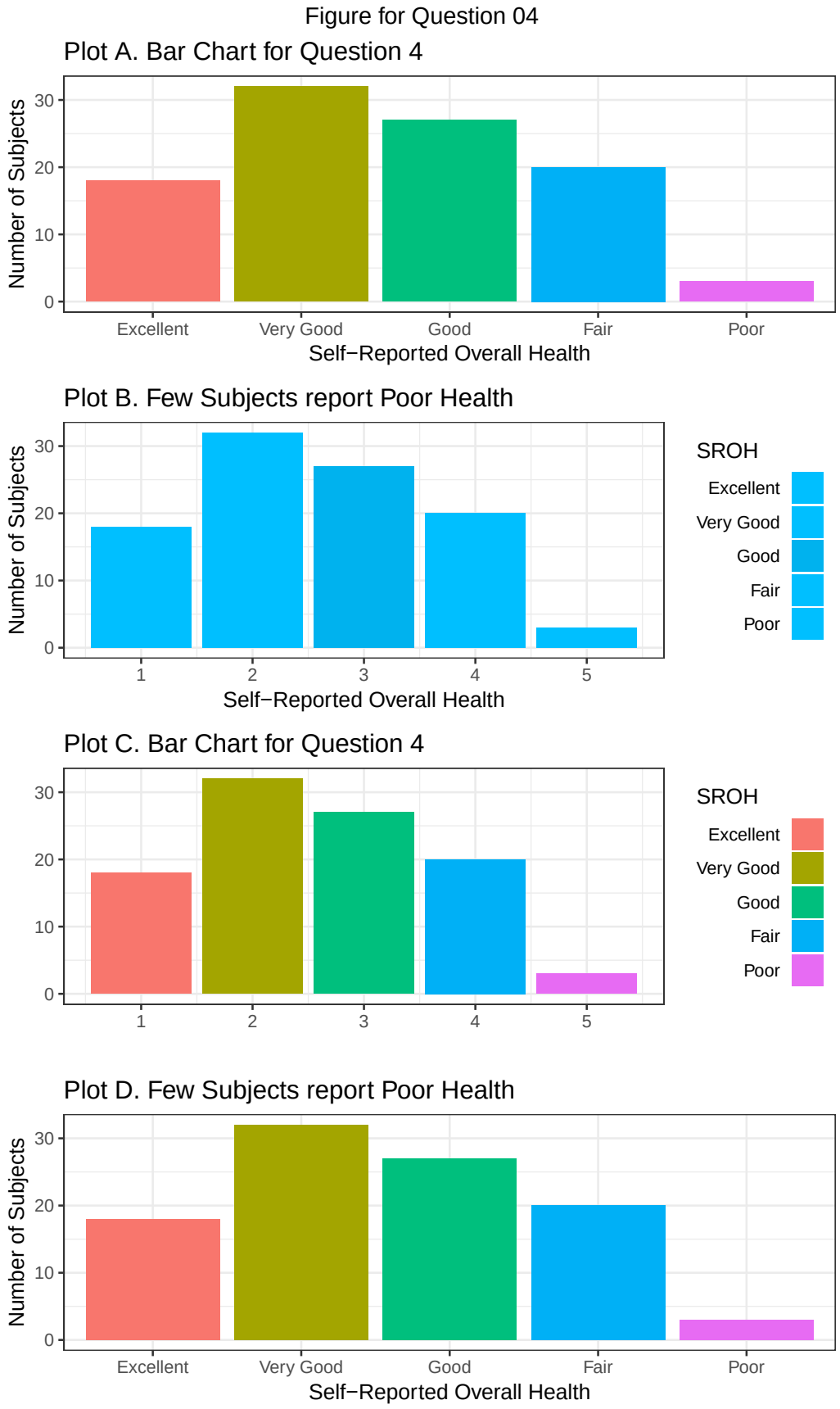
Results on Q03, worth 2 points.

- At least 46/51 (90% or more) gave the correct response. Aw!
- There was no partial credit for this question.
- The most common incorrect response was `b`.

Question 04

According to Jeff Leek in *The Elements of Data Analytic Style*, most of the following plots include something that should be **AVOIDED** in creating an effective visualization. One of the four plots does not include a problem of this sort. Please identify the good plot - the one that avoids Jeff's pitfalls.

Figure for Question 04



Answer for Question 04 is d

See Chapter 11 of Leek's *The Elements of Data Analytic Style*.

- Plot A is problematic because it uses a figure title that specifies the type of plot used, without describing the result.
- Plot B is problematic because it uses essentially indistinguishable colors for the fill in the bars, and doesn't explain the coding for Self-Reported Overall Health well unless you can distinguish these colors.
- Plot C is problematic because it has a poor title, and because it doesn't have labels on either the X or Y axis, and because it uses an unnecessary legend (the information on SROH should be incorporated into the labels on the X axis)
- Plot D is essentially reasonable. It is the best of these four plots.

Results on Q04, worth 2 points.

- At least 46/51 (90% or more) gave the correct response. Priceless!
- There was no partial credit for this question.
- The most common incorrect response was `a`.

Question 05

Passive exposure to environmental tobacco smoke has been associated with growth suppression and an increased frequency of respiratory tract infections in normal children. A study reported by B.K. Rubin in the New England Journal of Medicine (Sept 20 1990: "Exposure of children with cystic fibrosis to environmental tobacco smoke") looked at whether this association was more pronounced in children with cystic fibrosis. In a follow-up study, a new set of researchers measured a new set of 30 children, gathering each child's weight percentile and the number of cigarettes smoked per day in the child's home. For the 30 children in the new study, the Pearson correlation coefficient between weight percentile and cigarettes smoked was reported as r = -0.6. In interpreting the results in the responses below, the slope refers to the slope of a regression model predicting weight percentile using cigarettes smoked in the home for the 30 children. Which of the following interpretations of this result is most correct?

a. The slope will be negative, and the model will account for less than one-quarter of the variation in weight percentiles.
b. The slope will be positive, and the model will account for less than one-quarter of the variation in weight percentiles.
c. The slope will be negative, and the model will account for between 25% and 49% of the variation in weight percentiles.
d. The slope will be positive, and the model will account for between 25% and 49% of the variation in weight percentiles.
e. The slope will be negative, and the model will account for at least half of the variation in weight percentiles.
f. The slope will be positive, and the model will account for at least half of the variation in weight percentiles.
g. None of these interpretations are correct.

Answer for Question 05 is c

If r = -0.6, then r^2 will be 36% (so the model accounts for 36% of variation), and the slope will be negative.

Results on Q05, worth 2.5 points.

- 44/51 (86%) gave the correct response.
- The most common incorrect response was `e`.
- There was no partial credit for this question.

Question 06

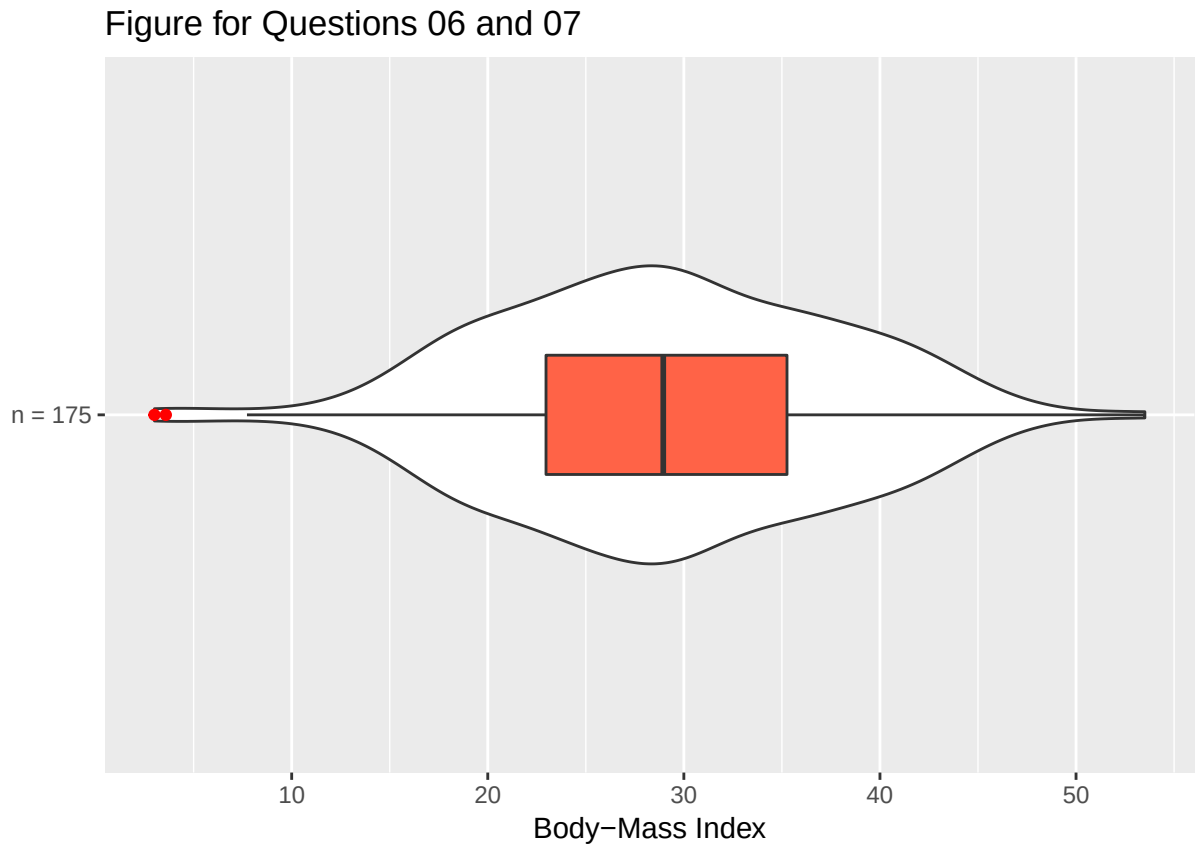
The Figure for Questions 06 and 07 displays the LDL cholesterol level (in mg/dl) for 175 adults who suffer from rheumatoid arthritis. The mean BMI in the sample is 29.2, the standard deviation is 8.7 and there are no missing values. Which of the following statements are true? (Check all that apply.)

The response options for Question 06 are:

a. The distribution is substantially left-skewed and cannot be approximated well with a symmetric model.
b. The median is about 34.
c. The IQR is about 12.
d. The distribution has too many outliers to be approximated well with a Normal model.
e. None of these statements are true.

Figure for Questions 06 and 07

```
set.seed(432201806)
temp <- rnorm(175, mean = 29, sd = 8)
dat06 <- data_frame(pt.id = 1:175, bmi = round(temp,2)-0.2)

ggplot(dat06, aes(x = "n = 175", y = bmi)) +
    geom_violin(width = 0.5) +
    geom_boxplot(fill = "tomato", width = 0.2, outlier.color = "red") +
    labs(x = "", y = "Body-Mass Index",
         title = "Figure for Questions 06 and 07") +
    coord_flip()
```



Answer for Question 06 is c, and only c

- The median is about 29, not 34.
- The data are not particularly asymmetric.
- There are 2 outliers in 175 observations, which is not in and of itself sufficient to render a Normal model impractical.

Here are the `favstats` results for these data.

```
mosaic::favstats(~ bmi, data = dat06)
```

```
 min    Q1 median    Q3   max     mean       sd   n missing
   3 22.97  28.95 35.26 53.51 28.99646 8.700906 175       0
```

Results on Q06, worth 2 points.

- At least 46/51 (90% or more) gave the correct response. Yay!
- The most common incorrect response was to include `a`, or `d` along with `c`.
- **Partial credit**: I gave you 1 point if you selected `c` along with something else.

Question 07

Adults with a BMI value of 25 or higher are classified as overweight. Based on the Figure for Questions 06 and 07, how many of the 175 patients would qualify as overweight by this standard?

a. Fewer than 43 patients.
b. Between 43 and 86 patients.
c. Exactly 87 patients.
d. Between 88 and 133 patients.
e. More than 133 patients.
f. There is insufficient information to answer the question.

Answer for Question 07 is d

Clearly 25 is below the median (so more than 50% of the 175 patients can have BMI 25 or more) but a good deal more than the 25th percentile (so at least 25% of the 175 patients have BMI below 25). So the correct answer is somewhere above 50% and below 75% of the 175 patients.

- Since 50% of 175 is 87.5 and 75% of 175 is 131.25, it looks like the right answer is between those values. That's response d.

Results on Q07, worth 3 points.

- 43/51 (84%) gave the correct response.
- The most common incorrect response was `e` but `f` was also popular.
- There was no partial credit for this question.

Question 08

The data set shown above, which is used in Questions 08, 09 and 10, shows arm and nose lengths of 20 women in a statistics class, and the ratio of arm to nose length for each. Rounded to one decimal place, what is the mean arm/nose ratio?

The Data Set

```
# A tibble: 20 x 4
   student ANratio   arm  nose
   <chr>     <dbl> <dbl> <dbl>
 1 Akari      15.2  66.8   4.4
 2 Beth       15.7  67.5   4.3
 3 Carol      15.5  68.1   4.4
 4 Donna      16.3  68.4   4.2
 5 Early      15.5  72.9   4.7
 6 Feng       15.8  72.8   4.6
 7 Grace      15.4  72.4   4.7
 8 Hanna      15.4  66.1   4.3
 9 Ione       14.9  65.6   4.4
10 Julie      13.5  64.6   4.8
11 Karen      15.1  63.5   4.2
12 Lin        15.7  70.5   4.5
13 Mary       15.1  63.3   4.2
14 Nancy      14.7  71.8   4.9
15 Olive      15.7  70.7   4.5
16 Paris      13.3  73.2   5.5
17 Ruo        16.4  68.7   4.2
18 Sara       13.7  71.3   5.2
19 Tilly      14.5  65.4   4.5
20 Vivi       14.6  62.8   4.3
```

Image of Data Set for Q08, Q09 and Q10

Student	Arm (in cm)	Nose (in cm)	Arm/Nose Ratio
Akari	66.8	4.4	15.2
Beth	67.5	4.3	15.7
Carol	68.1	4.4	15.5
Donna	68.4	4.2	16.3
Early	72.9	4.7	15.5
Feng	72.8	4.6	15.8
Grace	72.4	4.7	15.4
Hanna	66.1	4.3	15.4
Ione	65.6	4.4	14.9
Julie	64.6	4.8	13.5

Student	Arm (in cm)	Nose (in cm)	Arm/Nose Ratio
Karen	63.5	4.2	15.1
Lin	70.5	4.5	15.7
Mary	63.3	4.2	15.1
Nancy	71.8	4.9	14.7
Olive	70.7	4.5	15.7
Paris	73.2	5.5	13.3
Ruo	68.7	4.2	16.4
Sara	71.3	5.2	13.7
Tilly	65.4	4.5	14.5
Vivi	62.8	4.3	14.6

Answer for Question 08 is 15.1

If we add up the 20 arm-nose ratios in the data, we get 302. Dividing by 20, we get a mean of 15.1. Or, we could use R.

```
mosaic::favstats(~ ANratio, data = dat08)
```

```
  min     Q1 median   Q3  max mean        sd  n missing
 13.3 14.675   15.3 15.7 16.4 15.1 0.8516238 20       0
```

Results on Q08, worth 2 points.

- Everyone got this right. Mmm! This is shining!

Question 09

Refer to the data set for Q08, Q09 and Q10. What is the mode of the arm/nose ratios?

The response options for Q09 are:

a. 15.1
b. 15.3
c. 15.4
d. 15.7
e. None of the above.

Answer for Question 09 is d

15.7 is the mode. It occurs three times. No other value of arm-nose ratio occurs more than twice in our 20 observations. Arranging the data in order of this ratio can help us see this...

```
dat08 %>% select(student, ANratio) %>% arrange(ANratio)
```

```
# A tibble: 20 x 2
   student ANratio
   <chr>     <dbl>
 1 Paris      13.3
 2 Julie      13.5
 3 Sara       13.7
 4 Tilly      14.5
 5 Vivi       14.6
 6 Nancy      14.7
 7 Ione       14.9
 8 Karen      15.1
 9 Mary       15.1
10 Akari      15.2
11 Grace      15.4
12 Hanna      15.4
13 Carol      15.5
14 Early      15.5
15 Beth       15.7
16 Lin        15.7
17 Olive      15.7
18 Feng       15.8
19 Donna      16.3
20 Ruo        16.4
```

Results on Q09, worth 2 points.

- At least 46/51 (90% or more) gave the correct response. Phenomenal!
- The most common incorrect response was either `b` or `e`.
- There was no partial credit for this question.

Question 10

Refer to the data set for Questions 08-10. The Statue of Liberty's nose measures 4 feet, 6 inches, and her arm is 42 feet long. Calculate her arm/nose ratio, and use it to specify her Z score (# of standard deviations above or below the group mean) as compared to the 20 women described above. (Here's a hint to assist you with your calculations: The sample VARIANCE of the arm/nose ratio for the 20 women in our data set turns out to be 0.7253.) Your response should be the Z score for the Statue of Liberty, with an appropriate sign, rounded to one decimal place.

Answer for Question 10 is -6.8

To calculate the arm/nose ratio of the Statue of Liberty, we need to get her arm and nose lengths on the same scale. (Note that it doesn't have to be the same scale as was used for the women in the class, mathematically.) So, her arm length is 42 feet, and her nose length is 4.5 feet. Thus, the Statue of Liberty has arm/nose ratio of $42/(4.5) = 9.33$.

The mean of the 20 women is 15.1, and the standard deviation is the square root of the specified variance (0.7253) and $\sqrt{0.7253} = 0.8516$.

Thus, the Z score for the statue is

$$(9.33 - 15.1)/0.8516 = -6.775$$

And so Z = -6.8 for the Statue of Liberty after rounding to one decimal place.

Results on Q10, worth 3 points.

- 36/51 (71%) gave the correct response.
- The most common incorrect response was either `b` or `e`.
- **Partial Credit**: I gave 1 point on Question 10 for the responses -6.9 or -6.7.

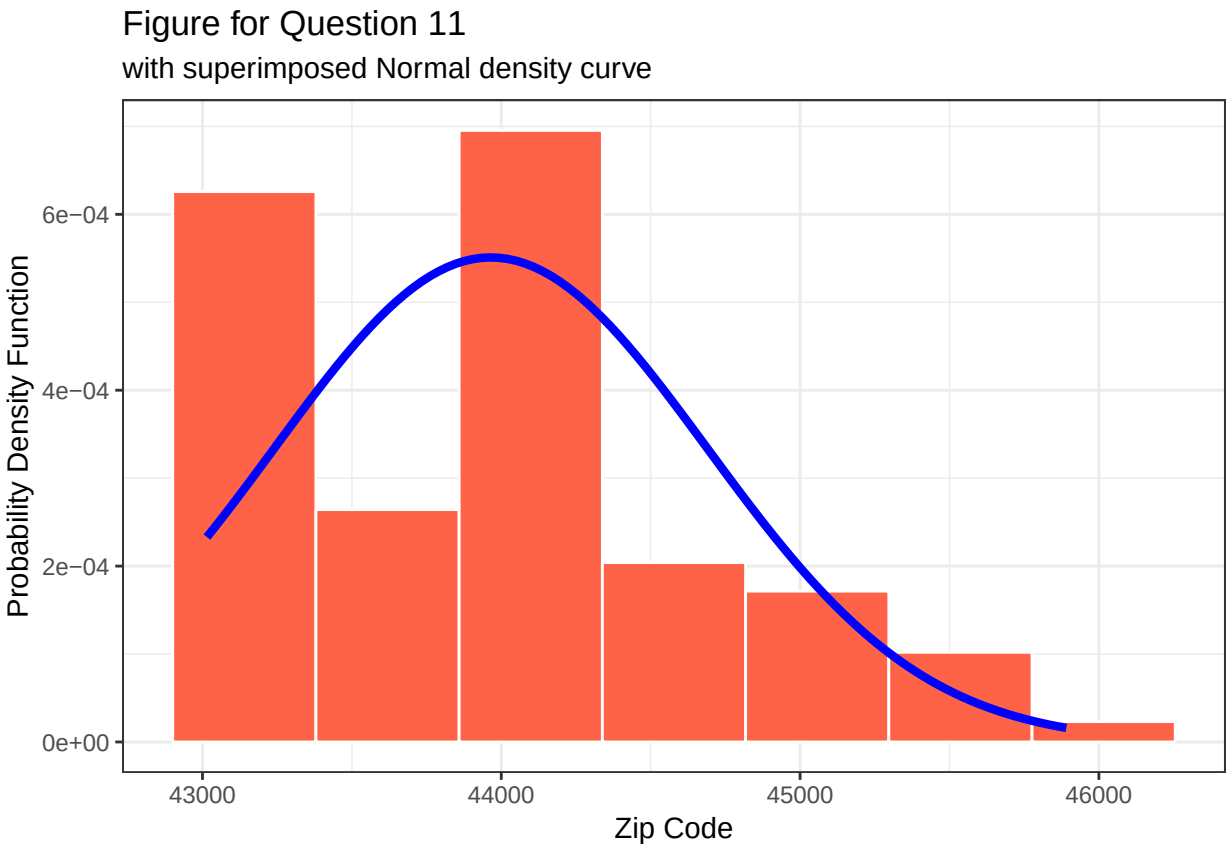
Question 11

The Figure for Question 11 shows the zip codes of the last 450 people from the state of Ohio to visit a web site providing information on purchasing insurance through the federal Health Insurance Marketplace. Which of the following summaries of these data would be most appropriate?

a. Mean
b. Median
c. IQR
d. Mode
e. It is impossible to tell from the information provided

Figure for Question 11

```
dat11 <- read.csv("data/dat11.csv") %>% tbl_df

ggplot(dat11, aes(x = zipcode)) +
    geom_histogram(aes(y = ..density..), bins = 7,
                   fill = "tomato", col = "white") +
    stat_function(fun = dnorm,
                  args = list(mean = mean(dat11$zipcode),
                              sd = sd(dat11$zipcode)),
                  lwd = 1.5, col = "blue") +
    labs(y = "Probability Density Function", x = "Zip Code",
         title = "Figure for Question 11",
         subtitle = "with superimposed Normal density curve") +
    theme_bw()
```



Answer for Question 11 is d

Zip codes are numbers, but they're not quantitative. Instead, they are nominal categorical data. Of these four choices, only a mode could possibly be relevant.

Results on Q11, worth 2 points.

- 21/51 (41%) gave the correct response.
- I expected this to be the bloodbath that it was, which was part of the reason I set it up for only two points.
- In terms of incorrect responses, people tried `b`, `e`, and `c`, all with about the same frequency. No one picked `a`, though.
- There was no partial credit for this question.

Question 12

Suppose you want to build a scatterplot to represent a linear regression model that predicts apgar5 (five minute APGAR scores) using apgar1 (one-minute APGAR scores) and the data for each variable are in the babydat data frame. Which of the following commands in R would be the most helpful in accomplishing this task?

a. ggplot(babydat, aes(x = apgar1, y = apgar5)) + geom_point() + geom_smooth(method = "lm")
b. ggplot(babydat, aes(x = apgar5, y = apgar1)) + geom_point() + geom_smooth(method = "loess")
c. ggplot(babydat, aes(x = apgar5, y = apgar1)) + geom_point() + geom_smooth(method = "lm")
d. ggplot(babydat, aes(x = apgar1, y = apgar5)) + geom_plot()
e. ggplot(babydat, aes(x = apgar1, y = apgar5)) + geom_smooth(method = "loess")
f. ggplot(babydat, aes(x = apgar5, y = apgar1)) + geom_plot()
g. None of these commands would be helpful

Answer for Question 12 is a

- To obtain a scatterplot, you use `geom_point`, not `geom_plot` (`geom_plot` isn't a thing) so we can eliminate d and f as options.
- The five-minute APGAR score needs to go on the vertical (y) axis because it is the outcome, and the one-minute APGAR score needs to go on the horizontal (x) axis because it is the predictor. So we need x = apgar1, and y = apgar5, so that eliminates b and c as options.
- Choice e shows a loess smooth, rather than the linear model we are trying to represent, and it doesn't show the actual points, so that eliminates e.
- That leaves choice a, which shows the linear model we are trying to represent. So that's the most helpful choice.

Results on Q12, worth 2 points.

- At least 46/51 (90% or more) gave the correct response. Heh!
- The most common incorrect response was `c`.
- There was no partial credit for this question.

Question 13

Consider the two histograms shown in the Figure for Question 13. On the left, we show the original data set, in a red color. On the right, we show the natural logarithm of the data, in a blue color. Assuming you are unsatisfied with assuming a Normal distribution for each of these expressions of the data, what transformation would the ladder of power transformations recommend next, in an effort to re-express the data in a form that could be modeled using a Normal distribution?

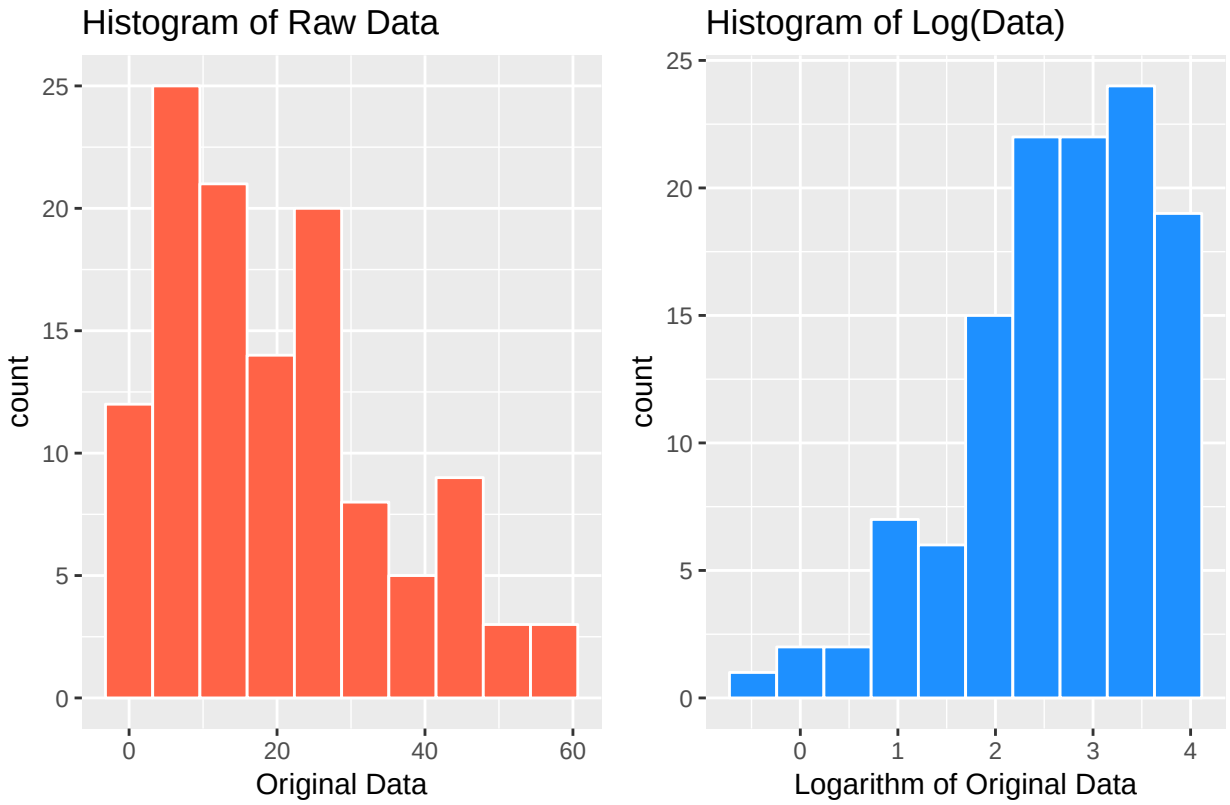
The response options for Question 13 are:

a. The square of the data
b. The square root of the data
c. The inverse of the data
d. The base 10 logarithm of the data
e. It is impossible to tell from the information provided

Figure for Question 13

```
set.seed(431013)
temp <- rchisq(120, df = 100, ncp = 100)
dat13 <- data_frame(pt.id = c(1:120), orig = abs(200 - temp))

p1 <- ggplot(dat13, aes(x = orig)) +
    geom_histogram(bins = 10, col = "white", fill = "tomato") +
    labs(title = "Histogram of Raw Data",
         x = "Original Data")

p2 <- ggplot(dat13, aes(x = log(orig))) +
    geom_histogram(bins = 10, col = "white", fill = "dodgerblue") +
    labs(title = "Histogram of Log(Data)",
         x = "Logarithm of Original Data")

gridExtra::grid.arrange(p1, p2, nrow = 1, top = "Figure for Question 13")
```

Figure for Question 13



Answer for Question 13 is b

Since the raw data are right skewed, and the logged data are left skewed, something in between seems the best choice. On the ladder of power transformations, the square root (transformation using power p = 0.5) falls between the raw data (p = 1) and the log (p = 0). ## Results on Q13, worth 3 points.

- 36/51 (71%) gave the correct response.
- The most common incorrect response was `c`.
- There was no partial credit for this question.

Question 14

Consider the data frame for Question 14, shown above, which describes scores on the Epworth Sleepiness Scale (ess_total) for 21 subjects, sorted from low to high. It turns out subject X (and only subject X) was pregnant at the time when the score was derived. If subject X is removed from the data set, which of the following statements are true? (Check all that apply.)

a. If we remove subject X, the mean will increase.
b. If we remove subject X, the median will increase.
c. If we remove subject X, the standard deviation will increase.
d. None of these statements are true.

Data Frame for Question 14

Subject	A	B	C	D	E	F	G	H	J	K	L
ESS_Total	3	9	5	6	4	9	10	0	4	5	2
Subject	M	N	P	Q	R	T	V	W	X	Z	
ESS_Total	12	0	7	11	1	18	16	12	22	8	

Answer for Question 14 is d and only d

None of these statements are true. In fact, in this case, the mean, median and standard deviation will actually decrease.

```
dat14 <- read.csv("data/dat14.csv") %>% tbl_df
```

Without subject X, the mean will decrease, since X has the maximum ESS score. The median will also decrease from 7 down to 6.5. The standard deviation will also decrease after X's removal.

Suppose we do the calculations...

- With subject X, we have mean 7.81, median 7 and sd 5.91:

```
mosaic::favstats(~ ess_total, data = dat14)
```

```
 min Q1 median Q3 max     mean       sd  n missing
   0  4      7 11  22 7.809524 5.912859 21       0
```

- Without subject X, we have mean 7.1, median 6.5 and sd 5.07:

```
mosaic::favstats(~ ess_total, data = dat14 %>% filter(subj_id != "X"))
```

```
 min   Q1 median    Q3 max mean       sd  n missing
   0 3.75    6.5 10.25  18  7.1 5.066921 20       0
```

Results on Q14, worth 2 points.

- At least 46/51 (90% or more) gave the correct response. Mmhm!
- The most common incorrect response was b.
- There was no partial credit for this question.

Question 15

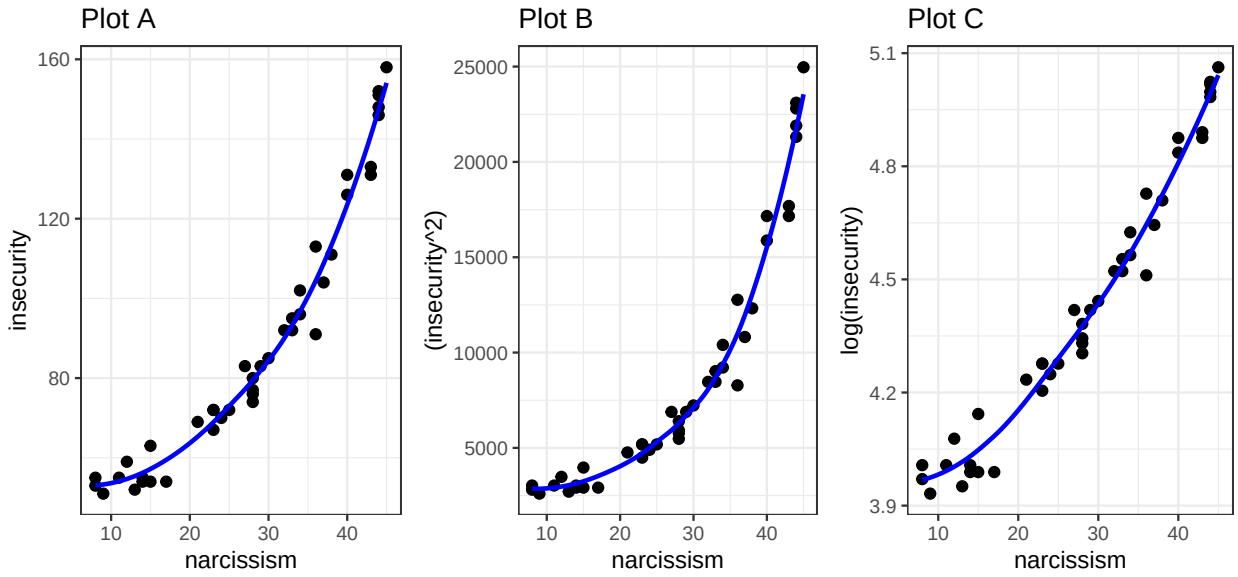
Question 15 describes the association of two variables, narcissism and insecurity, in a study of 42 subjects. The narcissism score is from a personality inventory designed to measure narcissism, and the insecurity score comes from an interview-based assessment of a subject's insecurity (and conversely, their self-esteem.) The principal investigator hypothesized prior to the study that the narcissism and insecurity scores would show an inverse (and potentially non-linear) relationship. The Figure for Question 15 shows three scatter plots with loess smooths, including the raw data for insecurity vs. narcissism, and then two different transformations (the square and the logarithm) of the insecurity scores, each plotted against the narcissism score. Which of the following is the analyst's BEST next step if her goal is to fit a linear model to the most appropriate transformation of the data she can find, using the ladder of power transformations?

a. Fit a linear model to predict the raw insecurity score based on the raw narcissism score.
b. Fit a linear model to predict the square of the insecurity score based on the raw narcissism score.
c. Build a new plot showing the cube of the insecurity score on the vertical (y) axis and the raw narcissism score on the horizontal (x) axis.
d. Build a new plot showing the inverse of the insecurity score on the vertical (y) axis and the raw narcissism score on the horizontal (x) axis.

Figure for Question 15

```
dat15 <- read.csv("data/dat15.csv") %>% tbl_df

p1 <- ggplot(dat15, aes(x = narcissism, y = insecurity)) +
    geom_point(size = 2) +
    geom_smooth(method = "loess", se = FALSE, col = "blue") +
    theme_bw() +
    labs(title = "Plot A")

p2 <- ggplot(dat15, aes(x = narcissism, y = (insecurity^2))) +
    geom_point(size = 2) +
    geom_smooth(method = "loess", se = FALSE, col = "blue") +
    theme_bw() +
    labs(title = "Plot B")

p3 <- ggplot(dat15, aes(x = narcissism, y = log(insecurity))) +
    geom_point(size = 2) +
    geom_smooth(method = "loess", se = FALSE, col = "blue") +
    theme_bw() +
    labs(title = "Plot C")

gridExtra::grid.arrange(p1, p2, p3, nrow = 1, top = "Figure for Question 15")
```

Figure for Question 15



Answer to Question 15 is d

- The raw data shows a rather pronounced curve that will not be picked up well by a straight line model, so choice a isn't the way to go.
- Taking the square of insecurity is clearly making the curve more pronounced in plot B, not less, so choice b isn't going to work, and also since this direction on the ladder moved us toward a more pronounced curve, choice c (fitting the cube) which goes even further in that direction, will also not be productive.
- The logarithm is a somewhat more promising transformation in that it straightens out the line a bit, but not really enough to justify a linear fit to the logarithm. An analyst trying to get the best possible transformation would like to consider moving a bit further down the ladder, past the logarithm to the inverse.

Results on Q15, worth 2 points.

- At least 46/51 (90% or more) gave the correct response. Polished!
- The most common incorrect response was `b` or `c`.
- There was no partial credit for this question.

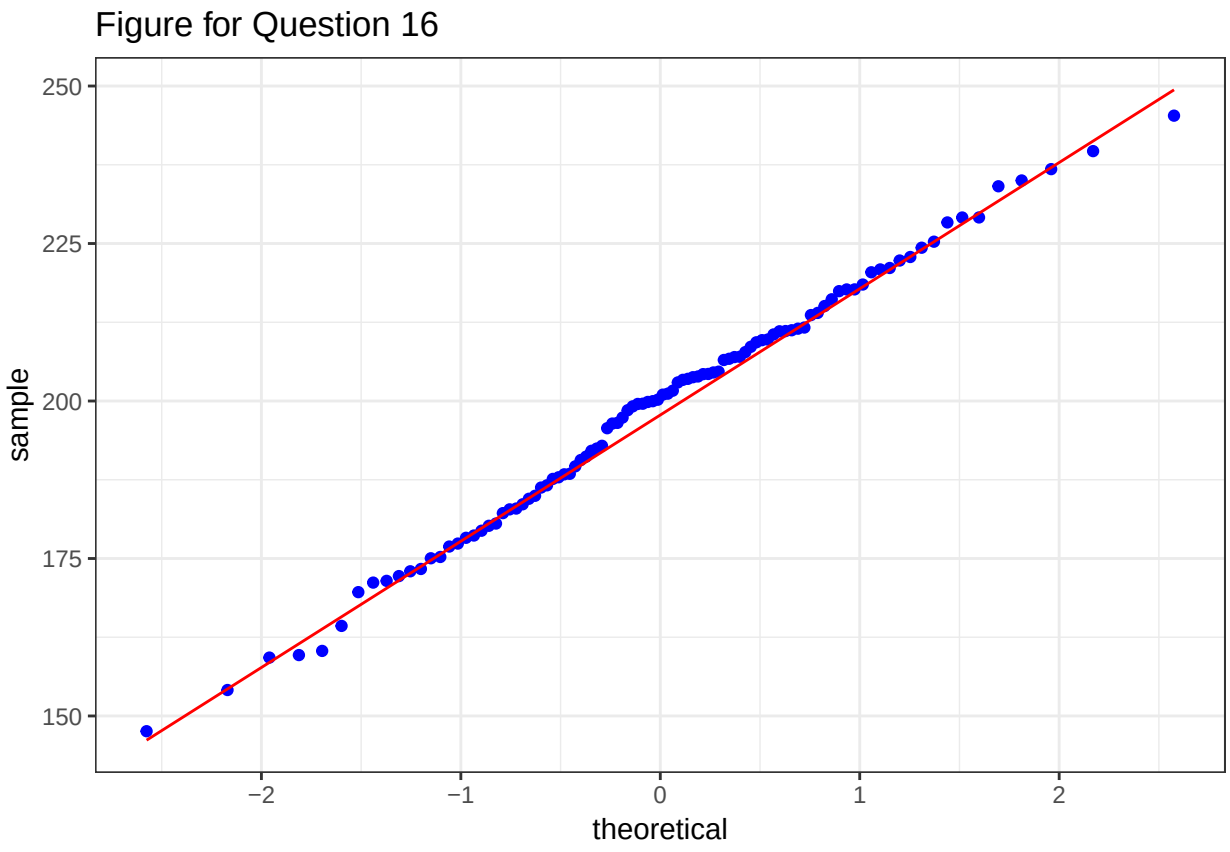
Question 16

The Figure for Question 16 is a Normal Q-Q plot of cholesterol levels (in mg/dl) for 100 adult American men. Which of the following statements best describes the distribution of the cholesterol levels?

a. Symmetric, but substantially outlier-prone in comparison to what we would expect from a Normal distribution.
b. Approximately Normally distributed, with a mean of approximately 250 mg/dl and a standard deviation of approximately 25 mg/dl.
c. Approximately Normally distributed, with a mean of approximately 200 mg/dl and a standard deviation of approximately 25 mg/dl.
d. Not approximately Normally distributed, but instead substantially left skewed.
e. Not approximately Normally distributed, but instead substantially right skewed.

Figure for Question 16

```
set.seed(43116)
item16 <- rnorm(100, mean=200, sd=23)
dat16 <- data_frame(item16)

ggplot(dat16, aes(sample = item16)) +
    geom_qq(col = "blue") +
    geom_qq_line(col = "red") +
    labs(title = "Figure for Question 16") +
    theme_bw()
```



Answer to Question 16 is c

- The data follow a straight line in the Normal Q-Q plot, and the center (mean) of the data is clearly near 200, with a standard deviation near 25, based on the Empirical Rule. There is no clear skew, nor are there substantial outliers, and the mean is clearly less than 250 mg/dl.
- If you viewed these data as symmetric, but outlier-prone, as a few people did, then you need to recalibrate your expectations for a Normal Q-Q plot.

Results on Q16, worth 2.5 points.

- 39/51 (76%) gave the correct response.
- By far, the most common incorrect response was `a`.
- There was no partial credit for this question.

Question 17

In a data frame called item17, you have a variable called apgar5 that contains scores on the APGAR scale at five minutes for 150 infants, although 4 of the values are listed as NA. You wish to obtain the standard deviation of the APGAR scores. If you need to know more about the APGAR score, visit https://goo.gl/9rxkVU. Your task is to mark the box next to EACH of the R commands listed below that produce the SAMPLE STANDARD DEVIATION of APGAR scores at five minutes for the 146 infants not marked as NA.

a. item17 %>% filter(complete.cases(apgar5)) %>% summarize(sd = sd(apgar5))
b. summary(item17)
c. item17 %>% select(complete.cases(apgar5)) %>% summarize(sd = sd(apgar5))
d. sd(apgar5)
e. item17 %>% summarize(sd(apgar5, na.rm = TRUE))
f. None of these will produce the correct value.

Answer to Question 17 is both a and e

- Statements a and e will produce the appropriate standard deviation.
- Statement b doesn't work because the `summary` function doesn't present the standard deviation.
- Statement c doesn't work because `select` is for picking columns (variables) rather than rows (observations) and you need `filter` to pick rows when using `complete.cases`.
- Statement d doesn't work because there are missing values in apgar5, so the result this gives is NA. You'd have to include na.rm=TRUE to make it work.

Results on Q17, worth 3 points.

- 35/51 (69%) gave the correct response.
- **Partial Credit**: If you identified `a, e` and nothing else, you got 3 points. If you identified both `a` and `e` but also included something else, then you got 2 points. If you got `a` or `e` by itself and didn't include anything else, I gave you 1 point.

Response	`a` and `e`	`a,e` and more	`a` or `e`, alone	Anything else
Count	35	8	6	2
Points	3	2	1	0

Question 18

A new sample of 300 subjects ages 35-59 from the NHANES data generates the Table for Question 18, which summarizes the relationship between the subject's Self-Reported Overall Health (Excellent, Vgood = "Very Good", Good, Fair or Poor) and whether or not they have ever tried marijuana (Yes/No). In this sample, which group is more likely to report their Self-Reported Overall Health as falling in one of the top three categories (Excellent, Very Good or Good)?

a. The "Yes" group, by more than three percentage points.
b. The "Yes" group, by 0.1 to 3 percentage points.
c. Neither group.
d. The "No" group, by 0.1 to 3 percentage points.
e. The "No" group, by more than three percentage points.
f. It is impossible to tell from the information provided.

Table for Question 18

```
library(NHANES)
dat18 <- NHANES %>% tbl_df

set.seed(431018)
dat18 <- dat18 %>%
  select(ID, Age, Marijuana, HealthGen) %>%
  filter(complete.cases(Age, Marijuana, HealthGen)) %>%
  filter(Age > 34, Age < 60) %>%
  sample_n(., size = 300) %>%
  droplevels()

knitr::kable(addmargins(table(dat18$Marijuana, dat18$HealthGen)))
```

	Excellent	Vgood	Good	Fair	Poor	Sum
No	9	41	55	20	2	127
Yes	22	60	62	26	3	173
Sum	31	101	117	46	5	300

Answer for Question 18 is b

Here are the health category percentages, within each marijuana group.0

```
knitr::kable(round(100*prop.table(table(dat18$Marijuana, dat18$HealthGen),1),1))
```

	Excellent	Vgood	Good	Fair	Poor
No	7.1	32.3	43.3	15.7	1.6
Yes	12.7	34.7	35.8	15.0	1.7

- So, in the "No" group, we have 7.1 + 32.3 + 43.3 = 82.7 percent in the three healthiest categories.
- In the "Yes" group, we have 12.7 + 34.7 + 35.8 = 83.2 percent in the three healthiest categories.
- So that is a difference of 0.5 percentage point, favoring the "Yes" group. That's choice **b**.

Results on Q18, worth 2 points.

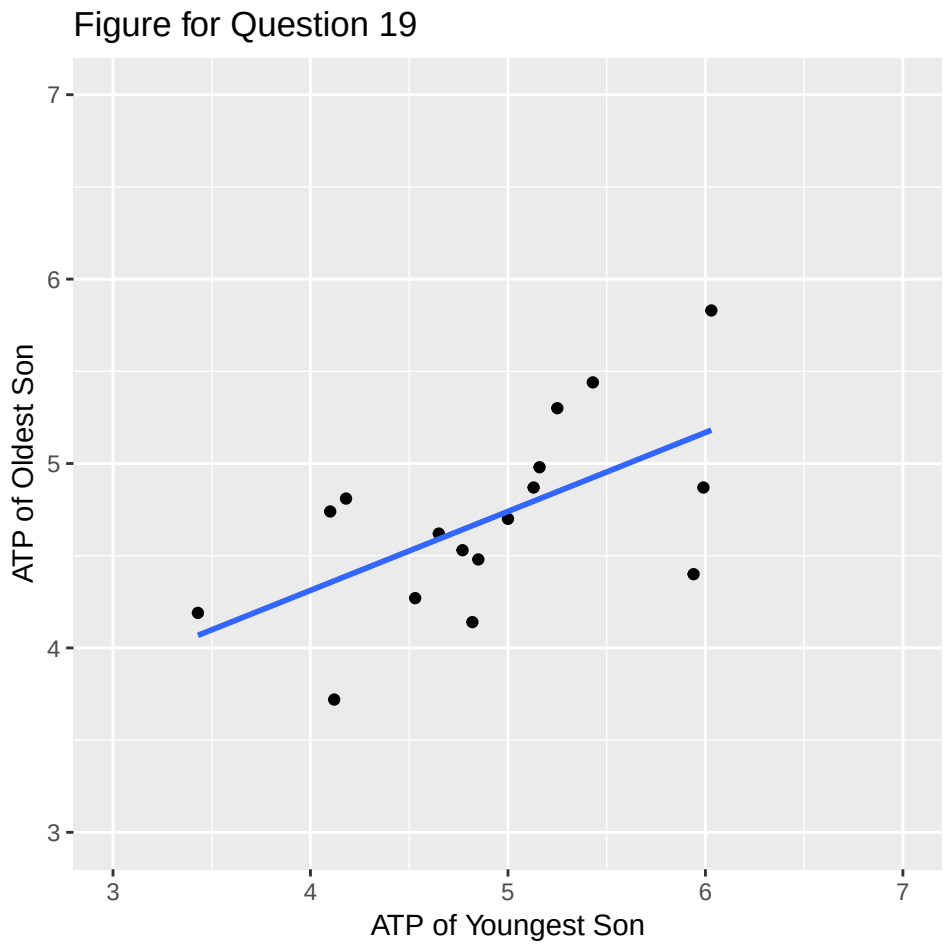
- 38/51 (75%) gave the correct response.
- The most common incorrect responses were `a` and `c`.
- There was no partial credit for this question.

Question 19

Dern and Wiorkowski (1969) collected data dealing with the erythrocyte adenosine triphosphate (ATP) levels in youngest and oldest sons in 17 families. The ATP level is an important measure of the ability of erythrocytes to transport oxygen in the blood. The Figure for Question 19 depicts the data for 17 pairs of brothers. Which of the following statements are true? (Check all that apply.)

a. The absolute value of the Pearson correlation is between 0 and 0.25.
b. The intercept of the regression line is less than zero.
c. The slope of the regression line is greater than zero.
d. None of these statements are true.

```
dat19 <- read.csv("data/dat19.csv") %>% tbl_df

ggplot(dat19, aes(x = young.atp, y = old.atp)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  xlim(3, 7) + ylim(3,7) +
  labs(x = "ATP of Youngest Son",
       y = "ATP of Oldest Son",
       title = "Figure for Question 19")
```



Answer for Question 19 is that c, and only c, is true

- Statement c is clearly true - the slope of the regression line is definitely positive. Higher levels of ATP of the youngest son are associated with higher ATP in the older son.
- Statement b requires a little thought, but extrapolating the line to see where it would cross the y-axis when the ATP of the youngest son is 0 suggests that the intercept is going to be somewhere between 2 and 3, in any case not negative, so Statement b is false. The actual regression line, as we see in the output below, is `old.atp` = 2.6 + 0.43 `young.atp`
- As for Statement a, The correlation is pretty strong here, in fact it turns out to be 0.6, as we see in the output below, and at any rate is much higher than 0.25. A correlation as low as 0.25 would indicate a very weak relationship, with points scattered far away from the straight line.

```
lm(old.atp ~ young.atp, data = dat19)
```

```
Call:
lm(formula = old.atp ~ young.atp, data = dat19)

Coefficients:
(Intercept)    young.atp
     2.5999       0.4281
```

```
cor(dat19 %>% select(old.atp, young.atp))
```

```
            old.atp young.atp
old.atp   1.0000000 0.5974007
young.atp 0.5974007 1.0000000
```

Results on Q19, worth 3 points.

- 39/51 (76%) gave the correct response (`c` and only `c`).
- **Partial Credit**: Everyone who got it wrong included other things besides `c` as true, with more picking `a` along with `c` than picked `b` along with `c`. At any rate, I gave all of those people (who picked `c` but also something else) one point.

Question 20

Again referring to the study discussed in Question 19, consider the ATP levels of the youngest brothers, for which the summary statistics shown in the Output for Question 20 are available. A non-parametric skew (skew1) calculation suggests that these data are. . .

a. Essentially symmetric
b. Seriously left-skewed
c. Seriously right-skewed
d. It is impossible to tell from the information provided

Output for Question 20

```
mosaic::favstats(~ young.atp, data = dat19)
```

```
  min   Q1 median   Q3  max     mean        sd  n missing
 3.43 4.53   4.85 5.25 6.03 4.904706 0.7173311 17       0
```

Answer for Question 20 is a.

The skew1 = (4.91 - 4.85)/0.72 = 0.08, which indicates no substantial skew. To show substantial skew, the nonparametric skew would have to be greater than 0.2 in absolute value. So, by this measure, the data are essentially symmetric.

Results on Q20, worth 2 points.

- At least 46/51 (90% or more) gave the correct response. Shining!
- About as many people chose `b` as `c`. No one chose `d`.
- There was no partial credit for this question.

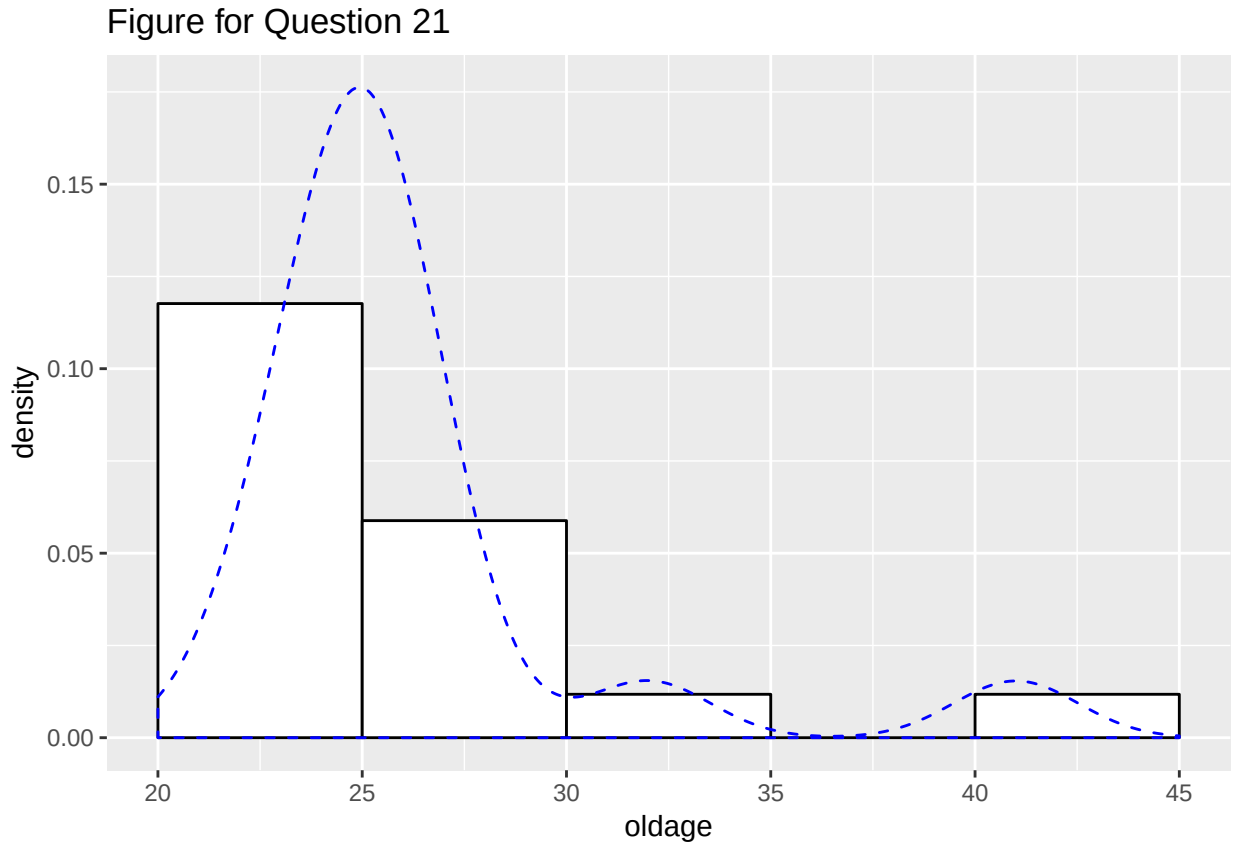
Question 21

Consider again the study described in Question 19, but now, we'll focus on the ages of the oldest sons. The figure below shows these ages (in years) for these 17 subjects, with a smooth density curve added. Which of the following statements are true? (Check all that apply.)

a. The mean of the ages is larger than the median age.
b. The ages are symmetric, showing no substantial skew.
c. The range of the data covers somewhere between 15 and 25 years.
d. None of these statements are true

Figure for Question 21

```
ggplot(dat19, aes(x = oldage, y = ..density..)) +
  geom_histogram(binwidth = 5, fill = "white", col = "black", center = 22.5) +
  geom_density(adjust = 2, lty = "dashed", col = "blue") +
  xlim(20,45) +
  labs(title = "Figure for Question 21")
```



Answer for Question 21 is both a and c.

- These are right-skewed data, according to the histogram, so statement a is true, and statement b is false.

- The data range from a bin marked 20-25 to a bin marked 40-45, so the range could be as small as 15 (40-25) and as large as 25 (45-20), so statement c is true, too.

Results on Q21, worth 2 points.

- 19/51 (37%) gave the correct response.
- Like Question 11, I expected this to be a problem for lots of people.
- In particular, I expected many people to answer `a` alone, and in fact, 20 people did.
- **Partial Credit**: I gave 1 point if you responded `a` alone or `c` alone.

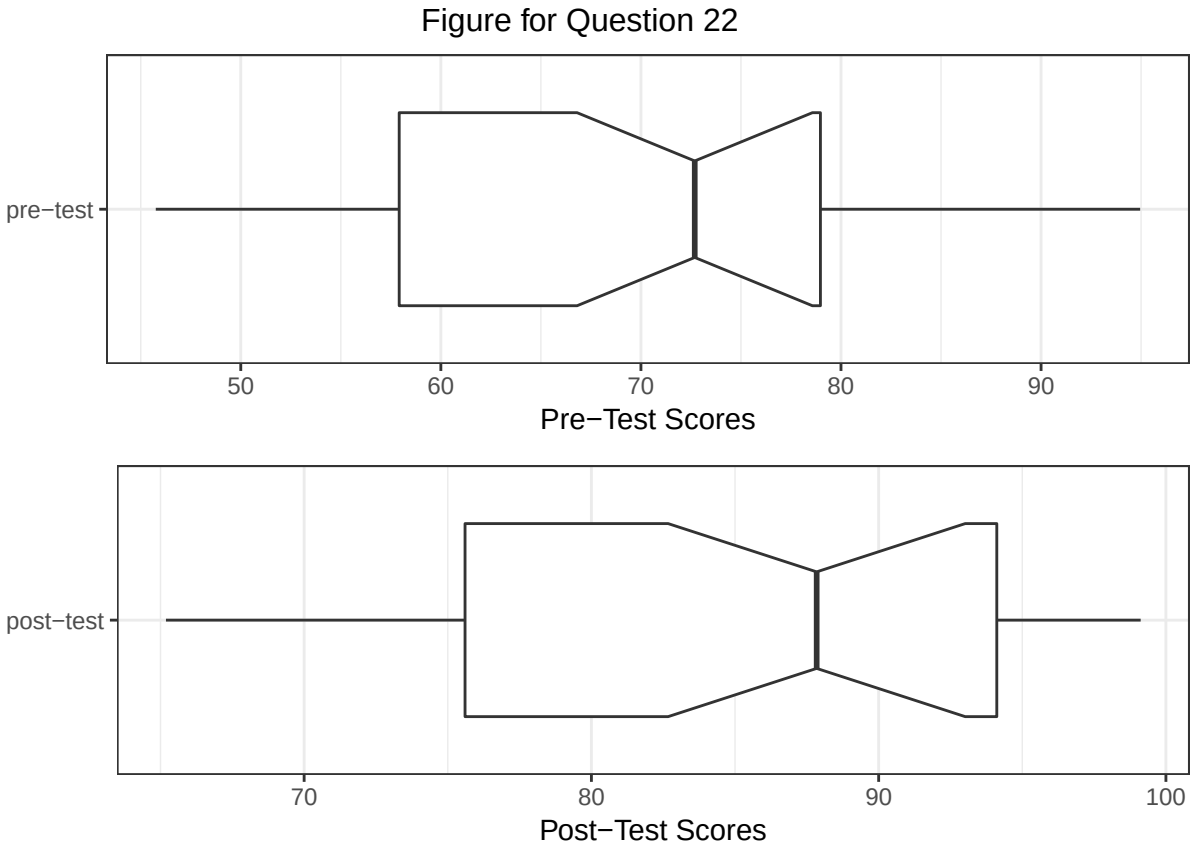
Question 22

Consider the pair of boxplots shown in the Figure for Question 22, which display pre-test results for a sample of 32 students in a statistics class, and then post-test results for a different sample of 32 other students in the same class. If you wanted to facilitate comparisons across the two plots in terms of center and spread of the distributions, what is the key change you would make to the Figure to accomplish this goal?

Figure for Question 22

```
set.seed(431022)
pre.test <- runif(32, min = 45, max = 95)
post.test <- runif(32, min = 65, max = 100)
dat22 <- data_frame(subj = 1:32, pre.test, post.test)
rm(pre.test, post.test)
```

```
plot22a <- ggplot(dat22, aes(x = "pre-test", y = pre.test)) +
    geom_boxplot(notch = TRUE) + coord_flip() + theme_bw() +
    labs(x = "", y = "Pre-Test Scores")

plot22b <- ggplot(dat22, aes(x = "post-test", y = post.test)) +
    geom_boxplot(notch = TRUE) + coord_flip() + theme_bw() +
    labs(x = "", y = "Post-Test Scores")

gridExtra::grid.arrange(plot22a, plot22b, ncol = 1,
                        top = "Figure for Question 22")
```



Answer for Question 22 is match the limits (scales) on the X-axis.

Rescale the plots so they have the same limits on the X-axis. I gave full credit to anything that would accomplish that end.

Results on Q22, worth 3 points.

- 42/51 (82%) gave a response I deemed correct.
- Answers I gave full credit to included:
  - Both boxplots should be displayed on the same horizontal axis scale.
  - Change the test score scale for one of the box plots so they match.
  - I would ensure that the scaling on the x-axis for both figures were the same.
- I didn't penalize you if you discussed matching the scales, but also were trying to do something else.
- But if you didn't try to match up the scales, and instead just suggested things like adding notches, or violins, or grid lines or sample means, you got no credit.
- **Partial Credit**: The relatively hard cases were the folks who said something about creating one plot, but were not specific enough about the need to align the scales. I gave these folks and those with similar responses, 1.5 points out of 3 for this question.
  - Plot the two box-plots on one single graph and compare both of them instead of two separate graphs
  - I would put both of the boxplots within the same chart. As in, put both into a comparison boxplot so that the differences are more pronounced.
  - In order to draw more accurate comparisons between these two boxplots, I would not have each box plot be it's own separate figure but instead have both boxplots be within the same figure separated by the factor of pre-test and post test. Then we would be able to make more accurate comparisons about the overlapping of the boxplot notches and data spread.
  - put them on the same plot
- I did give full credit to folks who talked about forming a single plot and also gave a reason involving the scales matching up, like ...
  - Compare the data on the same plot. The scale would be the same, therefore the data would be easier to interpret.
  - I would plot them in a single figure, so that each boxplot is using the same X-axis.

Question 23

How useful are exams given during the semester in predicting performance on the final? One class had three tests during the semester. A linear regression model was fit, yielding the (edited) output shown as the Output for Question 23. What percentage of the variation in final scores is accounted for by this regression model? Round your answer to one decimal place.

Output for Question 23

```
OUTPUT FOR QUESTION 23

Dependent Variable is Final

Predictor      Coeff   SE(Coeff)
Intercept     -32.84      14.00
Test1           0.35       0.23
Test2           0.70       0.21
Test3           0.39       0.22

Residual standard error = 13.46
Multiple R-squared = 64.1%
Adjusted R-squared = 60.2%
```

Answer for Question 23 is 64.1%

All you needed to do was provide the multiple R^2 value.

Results on Q23, worth 3 points.

- Only 37/51 (73%) gave the correct response, which was a bit disappointing.
- I thought in advance that this would be one of the easiest questions on the Quiz.
- The most common incorrect response was 60.2, but the adjusted R^2 is not what we need here.
- **Partial Credit**: One person rounded to 64%. I gave them 2 points.

Question 24

Again using the Output for Question 23, what is the predicted final score for a student with scores on Test1 of 68, on Test2 of 76, and on Test3 of 79? Round your final answer to zero decimal places.

Answer for Question 24 is 75.

```
-32.84 + 0.35*68 + 0.70*76 + 0.39*79
```

```
[1] 74.97
```

Our prediction is 75 after rounding to zero decimal places.

Results on Q24, worth 3 points.

- 43/51 (84%) gave the correct response. One person skipped the question. Don't do that.
- It was hard to see much of a pattern in the incorrect responses.
- **Partial Credit**: There were a couple of people who got 74 instead of 75. I gave them 1 point.

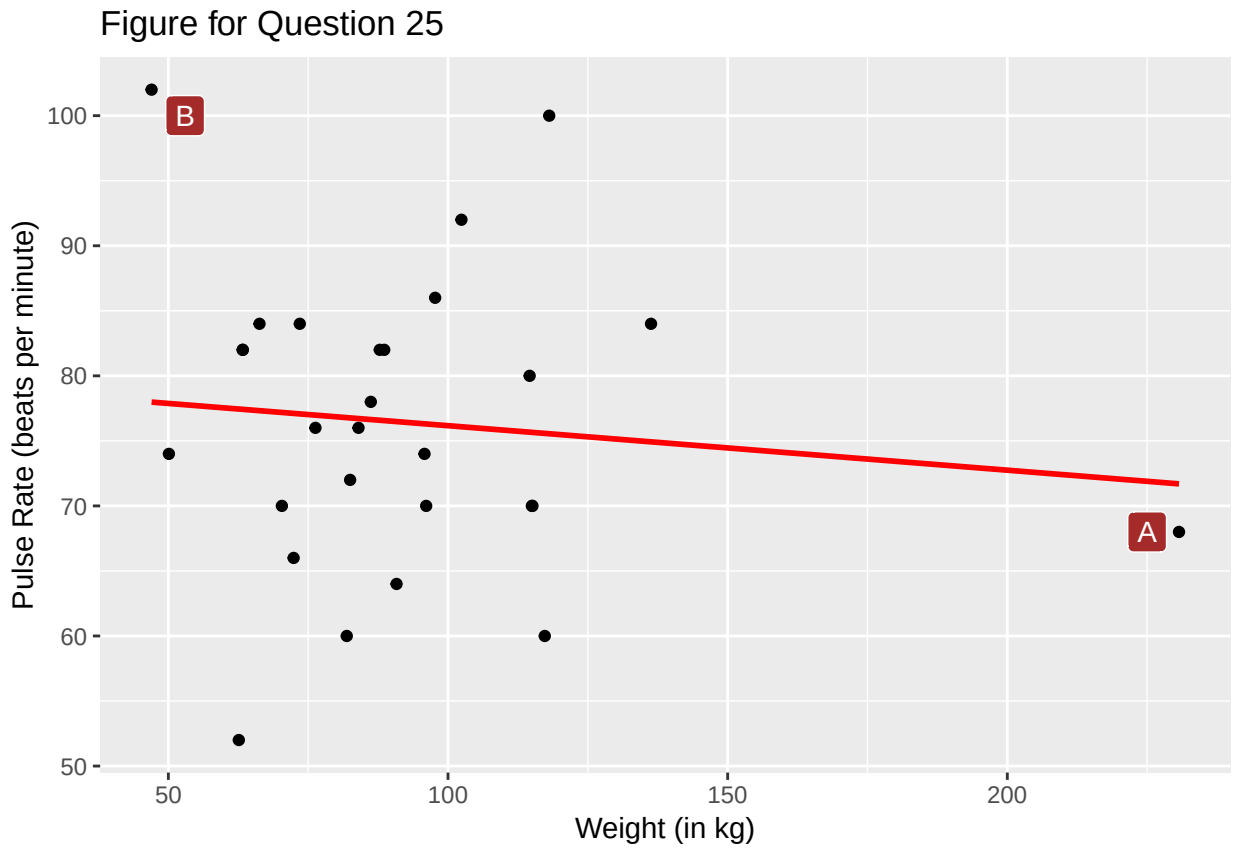
Question 25

Consider the scatterplot presented as the Figure for Question 25, which shows the relationship between an outcome of interest (pulse rate) and a predictor of interest (weight) for 28 subjects ages 31-59 from the NHANES data, including 2 subjects who are labeled A and B. Which of the following statements are true? (Check all that apply.)

a. The Pearson correlation coefficient including all points is positive.
b. The removal of point A will make the slope of the regression line increase.
c. The removal of both points (A and B) will make the slope of the regression line increase.
d. None of these statements are true

Figure for Question 25

```
library(NHANES)

dat25 <- NHANES %>% tbl_df

set.seed(919)
dat25 <- dat25 %>%
  select(ID, Age, Gender, Poverty, Weight, Pulse, TotChol, HealthGen) %>%
  filter(complete.cases(Age, Gender, Poverty, Weight, Pulse, TotChol, HealthGen)) %>%
  filter(Age > 30, Age < 60) %>%
  sample_n(., size = 28)

ggplot(dat25, aes(x = Weight, y = Pulse)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, col = "red") +
  annotate("label", x = 225, y = 68, label = "A",
           fill = "brown", col = "white") +
  annotate("label", x = 53, y = 100, label = "B",
           fill = "brown", col = "white") +
  labs(title = "Figure for Question 25",
       x = "Weight (in kg)",
       y = "Pulse Rate (beats per minute)")
```



Answer for Question 25 is both b and c.

- The slope of the regression line is negative when all points are included. Therefore, the Pearson correlation will also be negative. So statement a is false.
- Point A is dragging the line down on the right side. Removing it will therefore increase the slope (making it closer to 0 or even positive, it appears.) So statement b is true.
- Point B is pulling the line up on the left side, so removing it will also increase the slope. Removing *both* A and B will certainly increase the slope. So statement c is true.

Results on Q25, worth 2.5 points.

- 36/51 (71%) gave the correct response.
- The most common incorrect response was `c` alone.
- There was no partial credit for this question.

Question 26

Classify each of the following variables by their type.

The rows are:

a. Cause of death (for instance, homicide, heart failure, etc.)
b. Total body calcium of a patient with osteoporosis (to the nearest gram)
c. Days between attacks for a patient diagnosed with relapsing-remitting multiple sclerosis.
d. Province of residence for a group of Canadian citizens.
e. Self-reported amount of learning completed, based on a four item scale with the following responses for each item: didn't learn anything, learned a little bit, learned enough to be comfortable with the topic, learned a great deal.

The columns are:

Quantitative Ordinal categorical Nominal categorical It is impossible to tell

Answers for Question 26 are as follows:

- a. is Nominal categorical
- b. is Quantitative
- c. is Quantitative
- d. is Nominal categorical
- e. is Ordinal categorical

Results on Q26, worth 2.5 points total (0.5 per item).

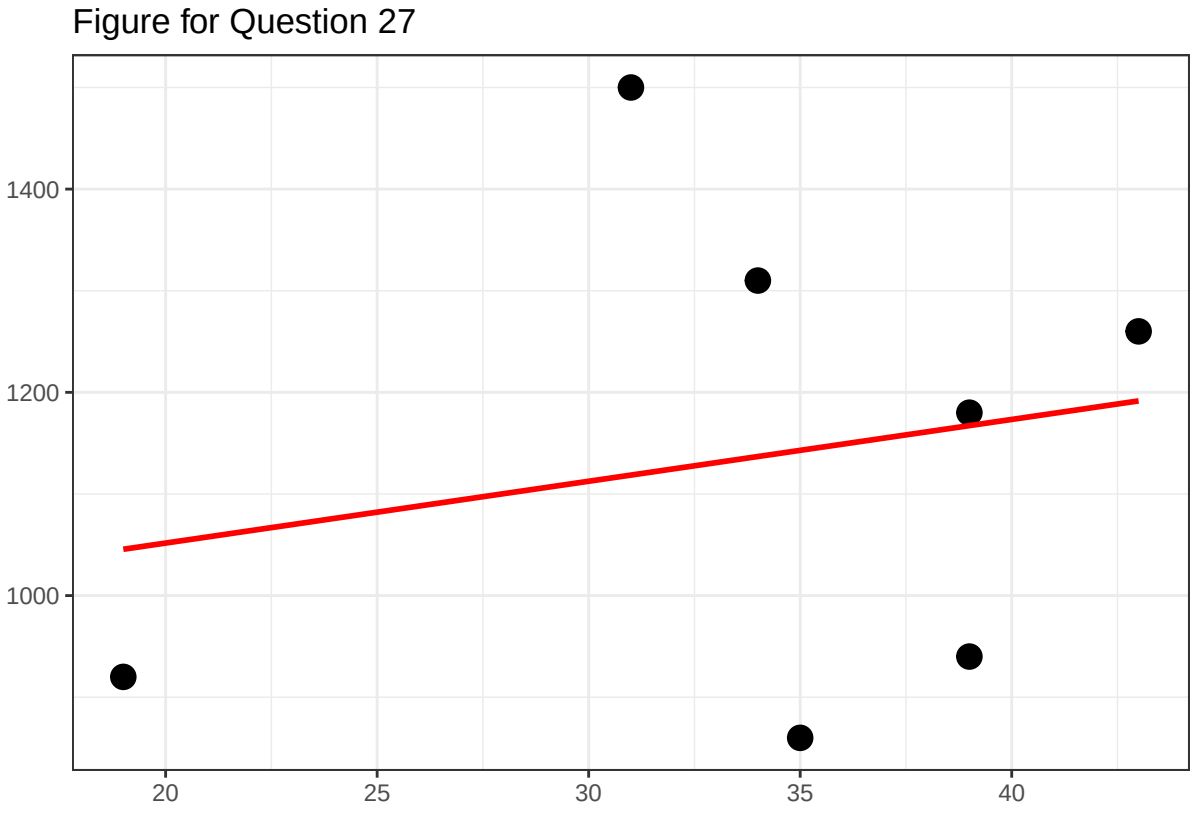
- Everyone got parts a, b, d and e right. Hooray! You all did that politely!
- In part `c`, more than 46/51 got it right. The other folks went for ordinal categorical, but it's a count of days. It has units. It's quantitative.
- You received 0.5 point for each correct response on Q26.

Question 27

Fast food is often high in both fat and sodium. But are the two related? The scatter plot shown in the Figure for Question 27 describes the fat (in g) and sodium (in mg) contents of several brands of hamburgers, and includes a linear model fit with geom_smooth, shown in red. In a sentence, what is the MOST IMPORTANT thing that should be done to improve the Figure?

Figure for Question 27

```
fat <- c(19, 31, 34, 35, 39, 39, 43)
sodium <- c(920, 1500, 1310, 860, 1180, 940, 1260)
dat27 <- data_frame(burger = 1:7, fat, sodium)
rm(fat, sodium)

ggplot(dat27, aes(x = fat, y = sodium)) +
    geom_point(size = 4) +
    geom_smooth(method = "lm", se = FALSE, col = "red") +
    labs(x = "", y = "", title = "Figure for Question 27") +
    theme_bw()
```



Answer for Question 27 is Add axis titles.

The correct response is to label the axes.

Results on Q27, worth 3 points.

- 45/51 (88%) gave a response I deemed correct.
- I gave full credit to anything that included "add axis labels" or something equivalent to that.
  - Some people also suggested things like adding a title for the graph, including units as part of the axis labels, and some people told me to use `labs` to do this. All OK, so long as you also asked for the axis labels. If you didn't, no credit.
- If you suggested adding a correlation coefficient as well as the axes, OK, but you had to call it by its correct name, and not something else, otherwise you would have lost a point.
- Those who didn't come up with this response usually focused on making some sort of transformation. If you suggested a transformation (for example) without also asking that we label the axes, you fell into my trap, and received no credit. But if you both asked for axis labels and a transformation, I let that slide and gave you full credit.

Question 28

Which set of data is more likely to have a bimodal shape?

a. Daily Cleveland, Ohio temperatures at noon in the month of July, 2018.
b. Daily Cleveland, Ohio temperatures at noon in 2018.
c. It is impossible to tell from the information provided.

Answer for Question 28 is b

Cleveland, like most cities far from the equator, has a period of time during the year where the temperature clusters around a fairly low value (winter) and another where it clusters around a fairly high value (summer).

```
dat29 <- read.csv("data/dat29.csv") %>% tbl_df
```

Results on Q28, worth 2 points.

- 31/51 (61%) gave the correct response, which was low, as expected.
- There was no partial credit for this question.

Response	a	b	c
Count	11	31	9

Output for Questions 29-32

Questions 29-32 make use of the `dat29` data that describe 40 patients with either aortic or mitral regurgitation who had heart surgery.

```
dat29
```

```
# A tibble: 40 x 7
      id ef.pre ef.post reg.type nyha  sbp.pre sbp.post
   <int>  <dbl>   <dbl> <fct>    <fct>   <int>    <int>
 1     1   0.51   0.36  mitral   II        150      120
 2     2   0.66   0.43  mitral   IV        125      124
 3     3   0.7    0.21  mitral   III       120      120
 4     4   0.39   0.26  aortic   IV        120      110
 5     5   0.41   0.17  aortic   I         150      110
 6     6   0.71   0.39  mitral   III       140      120
 7     7   0.68   0.77  mitral   II         90      110
 8     8   0.64   0.63  aortic   III       120      100
 9     9   0.56   0.5   aortic   I         160      110
10    10   0.56   0.290 aortic   IV        132      126
# ... with 30 more rows
```

The data are stored in the `dat29` tibble, as shown. The variables are:

- `id` = subject ID
- `ef.pre` = ejection fraction prior to surgery
- `ef.post` = ejection fraction after surgery
- `reg.type` = regurgitation type, either mitral or aortic
- `nyha` = NYHA class, an ordered four-category variable describing functional limitations
  - NYHA class levels are I, II, III and IV, with I indicating the least and IV indicating the most severe limitations
- `sbp.pre` = systolic blood pressure prior to surgery, in mm Hg.
- `sbp.post` = systolic blood pressure after surgery, in mm Hg.

Question 29

Write a single line of R code that will specify the coefficients of a linear regression model to predict systolic blood pressure after surgery on the basis of systolic blood pressure prior to surgery, using the dat29 tibble. Be sure that your code will work, and in particular, that you haven't spelled anything incorrectly.

Answer for Question 29 is a line of R code, like `lm(sbp.post ~ sbp.pre, data = dat29)`

Any code that would produce the estimated slope and intercept coefficients for the correct model is OK.

Good options include:

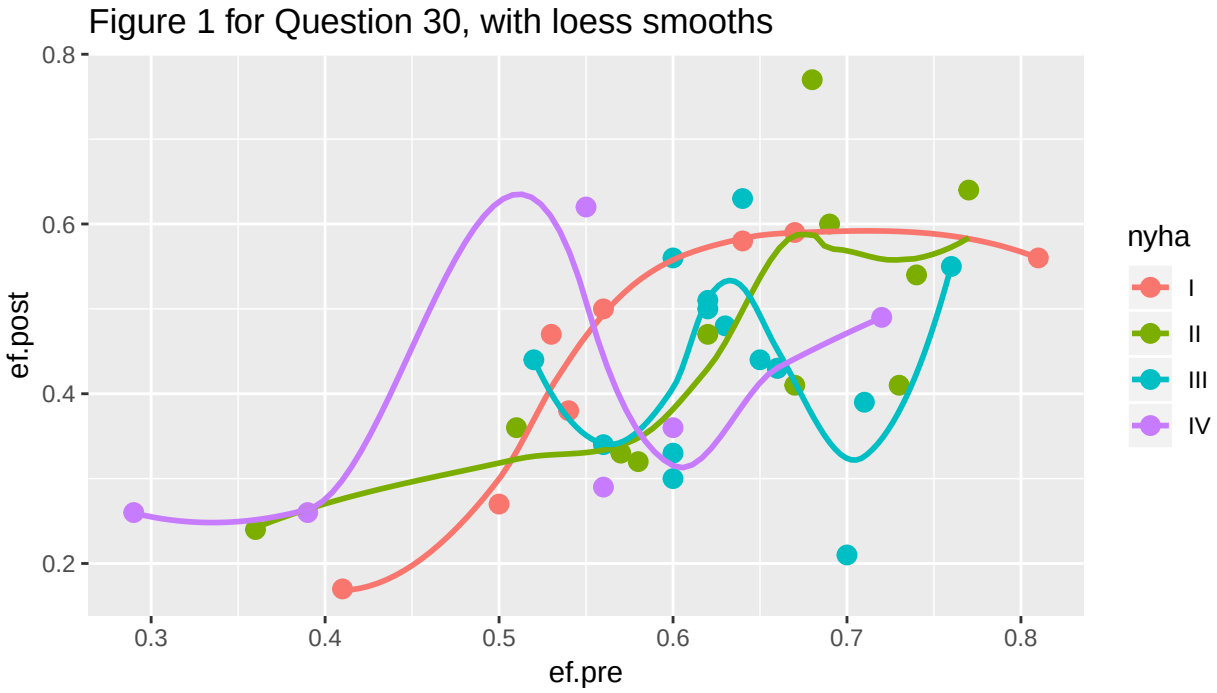
- `lm(sbp.post ~ sbp.pre, data = dat29)`
- `coef(lm(sbp.post ~ sbp.pre, data = dat29))`
- `summary(lm(sbp.post ~ sbp.pre, data = dat29))`
- `broom::tidy(lm(sbp.pre ~ sbp.post, data = dat29))`

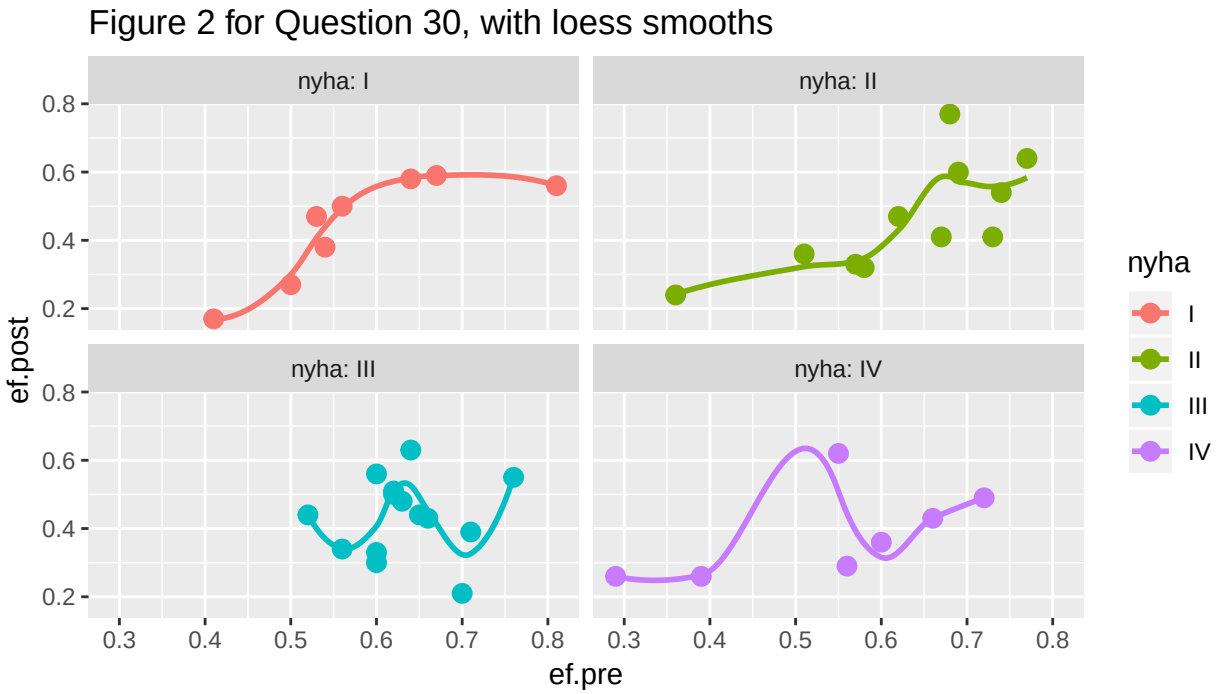
Results on Q29, worth 3 points.

- 34/51 (67%) gave a correct response. One person left it blank. Don't do that.
- **Partial Credit**:
  - Some people capitalized things like `Coef` or `DAT29`, which would cause an error in R. But if their code would have worked if they got the capitalization right, I gave them 1 point. Otherwise, zero.
  - Placing the line in an object, like `dat29_line<-lm(sbp.post~sbp.pre, data=dat29)` doesn't provide you with any results - it just stores them, so just 1 point for that.
  - Running something that doesn't actually produce the model's coefficients gets no credit.
  - Asking for the correlation would not have helped here.
  - Running a ggplot would not have been helpful here.
  - Including `scale`, as in `summary(lm(sbp.post~scale(sbp.pre), data= dat29))` doesn't get you the right model, so no points.

Question 30

This question also makes use of the Setup for Questions 29-32. Figure 1 for Question 30 and Figure 2 for Question 30 each show the same data, and the same loess smooths, using two different approaches. Please specify the one-line R command I added to Figure 1 in order to achieve Figure 2. . In your response, please ignore the fact that I also changed the title of the Figure.





Answer for Question 30 is `facet_wrap(~ nyha, labeller = "label_both")`

I gave credit, regardless of whether you included or excluded the + sign, so that either `facet_wrap(~ nyha, labeller = "label_both")` or `facet_wrap(~ nyha, labeller = "label_both") +` would be correct.

Results on Q30, worth 3 points.

- 4/51 (8%) gave a completely correct response, although I awarded enough partial credit that 61% of available points were in fact awarded. There were lots of problems.
- It's `nyha` and not `NYHA` and to R, that matters, so if you did that, you lost a point.
- Leaving out the tilde (~) cost you a point.
- Leaving out the labeller = "label_both" would also cost you a point. That's what happened to many, many people.
- You needed to get `nyha` right, using a different variable lost you all credit.
- You could have used `nrow = 2` or `ncol = 2` but if you didn't, you were still ok, since the machine will split four plots into two rows by default, so leaving that out doesn't lose any credit.
- Someone used `facet_wrap(~ dat29$nyha)` which wouldn't actually work. `ggplot2` specifies the data set elsewhere (in the `aes` section.)
- Just saying `facet_wrap` without anything else got you no credit.
- Including the entire ggplot command lost you one additional point.
- You could have also used `facet_grid` to get facets, but that would have produced a different result. Nonetheless, if you used `facet_grid(~ nyha)` or `facet_grid(nyha ~ .)`, I gave you credit for 2 points, and 3 if you had included the labeller piece.

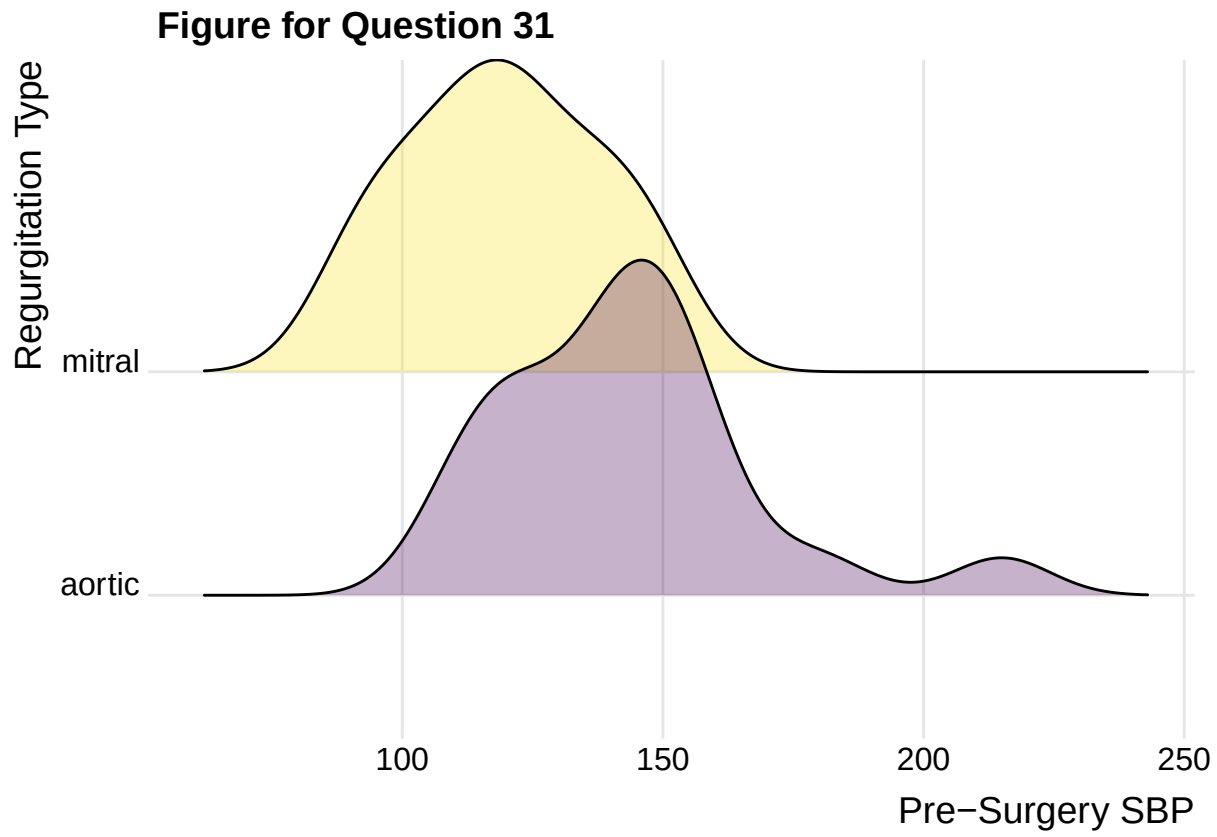
Question 31

This question also makes use of the Setup for Questions 29-32. The Figure for Question 31 shows the pre-surgery systolic blood pressures for the same patients for the two types of regurgitation, using what is called a ridgeline plot, built using the `ggridges` package. Which regurgitation type displays a larger sample mean pre-surgery systolic blood pressure?

a. Aortic regurgitation
b. Mitral regurgitation
c. It is impossible to tell from the information provided.

Figure for Question 31

```
library(ggridges)

ggplot(dat29, aes(x = sbp.pre, y = reg.type, fill = reg.type)) +
    geom_density_ridges(alpha = 0.3, scale = 1.5) +
    scale_fill_viridis_d() +
    guides(fill = FALSE) +
    labs(title = "Figure for Question 31",
         x = "Pre-Surgery SBP",
         y = "Regurgitation Type") +
    theme_ridges()
```



Answer for Question 31 is a

The center of the distribution of the aortic regurgitation data is clearly to the right of the mean in the mitral regurgitation data and thus the mean in the aortic group will be larger, according to the ridgeline plot.

Results on Q31, worth 2 points.

- At least 46/51 (90% or more) gave the correct response. Flawless!
- There was no partial credit for this question.

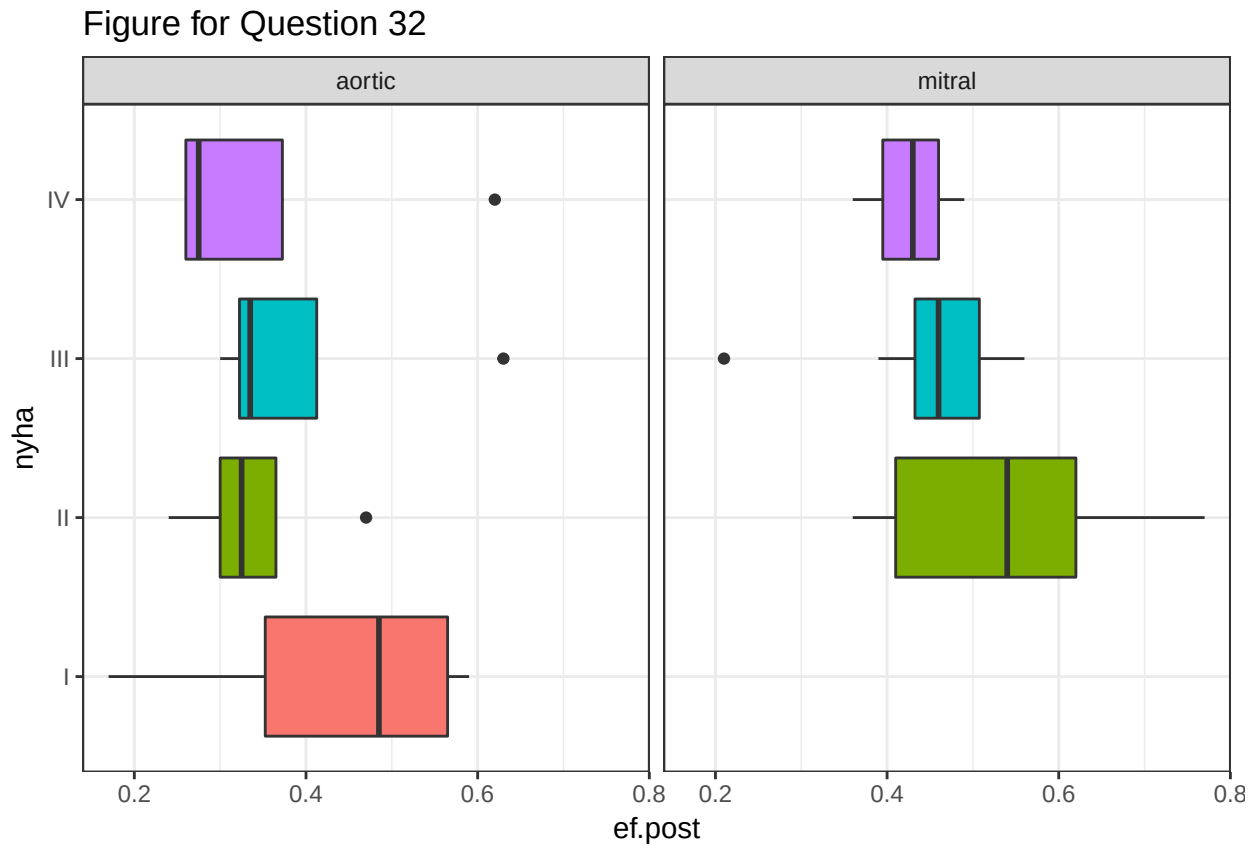
Question 32

This question is the final one which makes use of the Setup for Questions 29-32. The Figure for Question 32 shows the ejection fraction after surgery for the 40 patients, and the complete code used to develop the Figure is also provided. Which of the following lines of R code would best inform you as to why there are only seven boxplots in the Figure for Question 32 rather than eight?

a. drop_na
b. facet_grid(~ reg.type, labeller = "label_both")
c. summary(dat29)
d. dat29 %>% count(reg.type, nyha)
e. dat29 %>% group_by(nyha) %>% summarize(reg.type)
f. None of these would be useful.

Figure for Question 32

```
ggplot(dat29, aes(x = nyha, y = ef.post, fill = nyha)) +
    geom_boxplot() +
    facet_wrap(~ reg.type) +
    coord_flip() +
    guides(fill = FALSE) +
    labs(title = "Figure for Question 32") +
    theme_bw()
```



Answer for Question 32 is d

As we can see from the results of applying the code in **d** below, there are no subjects in the nyha I group who had mitral regurgitation in the data set. That's why no data are plotted in the bottom right of the Figure for Question 32. None of the other codes would provide us with this information, although there are other ways we could have used to figure this out.

```
dat29 %>% count(reg.type, nyha)
```

```
# A tibble: 7 x 3
  reg.type nyha      n
  <fct>    <fct> <int>
1 aortic   I         8
2 aortic   II        4
3 aortic   III       4
4 aortic   IV        4
5 mitral   II        7
6 mitral   III      10
7 mitral   IV        3
```

Results on Q32, worth 2.5 points.

- Only 19/51 (37%) gave the correct response. I had hoped this would go better, but I knew it was hard.
- The most common incorrect responses were `a` and `e` but `c` and `f` were also selected by multiple people.
- There was no partial credit for this question.

Question 33

Consider the `starwars` data set that is part of the `dplyr` package in the tidyverse. How many of the characters listed in that data set are of the Human species and have blue eye_color? (Note that we ask for blue, specifically, here, and not other colors that might be various shades of blue.)

Answer is 12.

```
humanblue <- starwars %>% filter(eye_color == "blue" & species == "Human")

count(humanblue)
```

```
# A tibble: 1 x 1
      n
  <int>
1    12
```

Results on Q33, worth 3 points.

- 44/51 (86%) gave the correct response.
- The most common incorrect answer was 19, and I'm sure there's a reason for that, probably people looking for additional shades of blue.
- There was no partial credit for this question.

Question 34

Suppose you built a subset of the `starwars` data called `humanblue` which consists only of the characters who are Human with blue eyes, and that you now want to obtain the median of their mass in kilograms, among those subjects who have a mass recorded. Which of the following lines of R code would do that? Check all that apply.

a. summary(humanblue %>% select(mass))
b. humanblue %>% filter(complete.cases(mass)) %>% summarize(quantile(mass, probs = 0.5))
c. humanblue %>% summarize(median(mass, na.rm = TRUE))
d. humanblue %>% filter(complete.cases(mass)) %>% summarize(median(mass))
e. mosaic::favstats(~ mass, data = humanblue)
f. None of these.

Answer to Question 34 is that all of them (a, b, c, d, and e) work.

```
summary(humanblue %>% select(mass))
```

```
      mass
 Min.   : 75.00
 1st Qu.: 78.00
 Median : 84.00
 Mean   : 90.57
 3rd Qu.: 99.50
 Max.   :120.00
 NA's   :5
```

```
humanblue %>% filter(complete.cases(mass)) %>% summarize(quantile(mass, probs = 0.5))
```

```
# A tibble: 1 x 1
  `quantile(mass, probs = 0.5)`
                          <dbl>
1                            84
```

```
humanblue %>% summarize(median(mass, na.rm = TRUE))
```

```
# A tibble: 1 x 1
  `median(mass, na.rm = TRUE)`
                         <dbl>
1                           84
```

```
humanblue %>% filter(complete.cases(mass)) %>% summarize(median(mass))
```

```
# A tibble: 1 x 1
  `median(mass)`
           <dbl>
1             84
```

```
mosaic::favstats(~ mass, data = humanblue)
```

```
 min Q1 median   Q3 max     mean       sd n missing
  75 78     84 99.5 120 90.57143 17.55807 7       5
```

Results on Q34, worth 3 points.

- 31/51 (61%) gave the correct response, but I gave partial credit so that 71% of available points were awarded.
- The most common incorrect response was to leave out `a`, and maybe also `e`.
- **Partial credit**: I gave 2 points if you selected 4 of the five correct responses, and 1 if you selected 3 out of the 5.

Question 35

I produced the cross-tabulation shown in the Output for Question 35 using the complete `starwars` tibble available in the `tidyverse`. The magrittr and tidyverse packages were already loaded on my computer. Which of the following commands did I use?

a. table(gender, height > 170, data = starwars)
b. mosaic::favstats(~gender + height > 170, data = starwars)
c. starwars %$% table(gender, height>170)
d. starwars %>% select(gender, height) %>% table(height > 170)
e. starwars %$% filter(height > 170) %>% count(gender)
f. None of these would work.

Output for Question 35

```
starwars %$% table(gender, height > 170)
```

```
gender          FALSE TRUE
  female           12    5
  hermaphrodite     0    1
  male             12   47
  none              0    1
```

Answer for Question 35 is c

None of the other codes would produce this result. Some wouldn't produce anything but an error message.

Results on Q35, worth 3 points.

- 42/51 (82%) gave the correct response.
- The most common incorrect answers were `d` and `f`.
- There was no partial credit for this question.

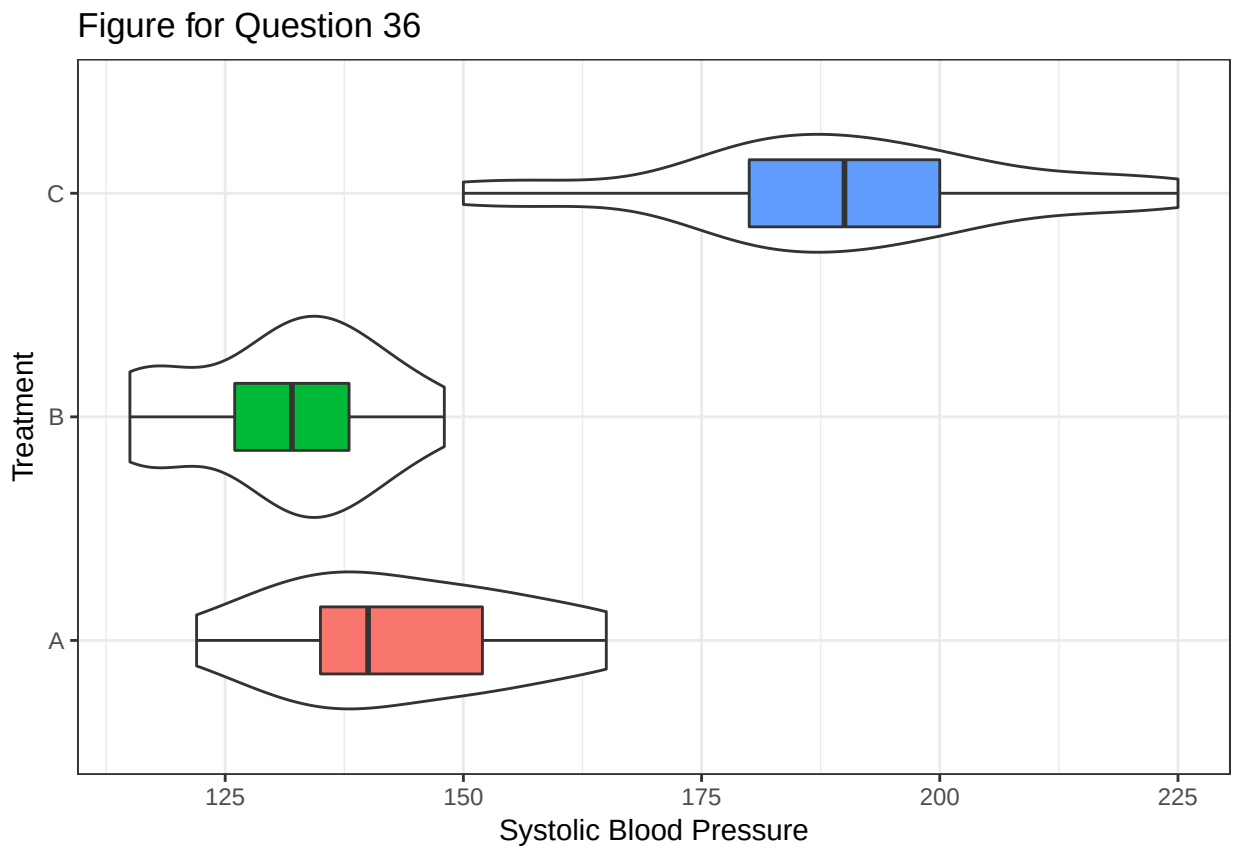
Question 36

Data from a paper by Vlachakis and Mendlowitz (1976) dealt with the treatment of essential hypertension ("essential" is a technical term here meaning the cause is unknown - a synonym is "idiopathic".) Seventeen patients received treatments C, A and B, where C = Control Period, A = Propranolol + Phenoxybenzamine and B = Propranolol + Phenoxybenzamine + Hydrochlorothiazide Each patient received C first, then either A or B, and then the remaining treatment. The data consist of systolic blood pressures under the three conditions. The Figure for Question 36 compares the results. Which of these three distributions, A, B or C, shows the largest spread?

a. Treatment A
b. Treatment B
c. Treatment C
d. It is impossible to tell from the information provided

Figure for Question 36

```
dat36 <- read.csv("data/dat36.csv") %>% tbl_df

ggplot(dat36, aes(x = treatment, y = sbp)) +
    geom_violin() +
    geom_boxplot(aes(fill = treatment), width = 0.3) +
    coord_flip() +
    theme_bw() +
    guides(fill = FALSE) +
    labs(y = "Systolic Blood Pressure", x = "Treatment",
         title = "Figure for Question 36")
```



Answer for Question 36 is c

This is evident from the whiskers, which show a much larger range in Treatment C results than in either of the other treatment groups.

Results on Q36, worth 2 points.

- At least 46/51 (90% or more) gave the correct response. Super-excellent!
- There was no partial credit for this question.

Question 37

There are four figures provided above as Figures A, B, C, and D for Question 37. Suppose you are using a subset of the `midwest` data from the `ggplot2` package. You are trying to determine for this subset whether or not a transformation of the outcome (specifically, taking the inverse of the outcome) is necessary to fit a linear regression model to describe the relationship between percollege (the predictor, specifically the percent college educated) and percbelowpoverty (the outcome, specifically the percent below the poverty level). Which of the Figures for Question 37 would be of the most help in assessing the impact of the inverse transformation on the quality of fit for a linear model?

a. Figure A
b. Figure B
c. Figure C
d. Figure D
e. They would all be equally useful

Figure A for Question 37

```
midwest %>% filter(state == "OH") %>%
    ggplot(., aes(x = percollege, y = percbelowpoverty)) +
    geom_point() +
    geom_smooth(method = "loess", col = "blue") +
    geom_smooth(method = "lm", se = FALSE, col = "red") +
    labs(title = "Figure A for Question 37",
         subtitle = "with linear and loess fits for percbelowpoverty by percollege")
```

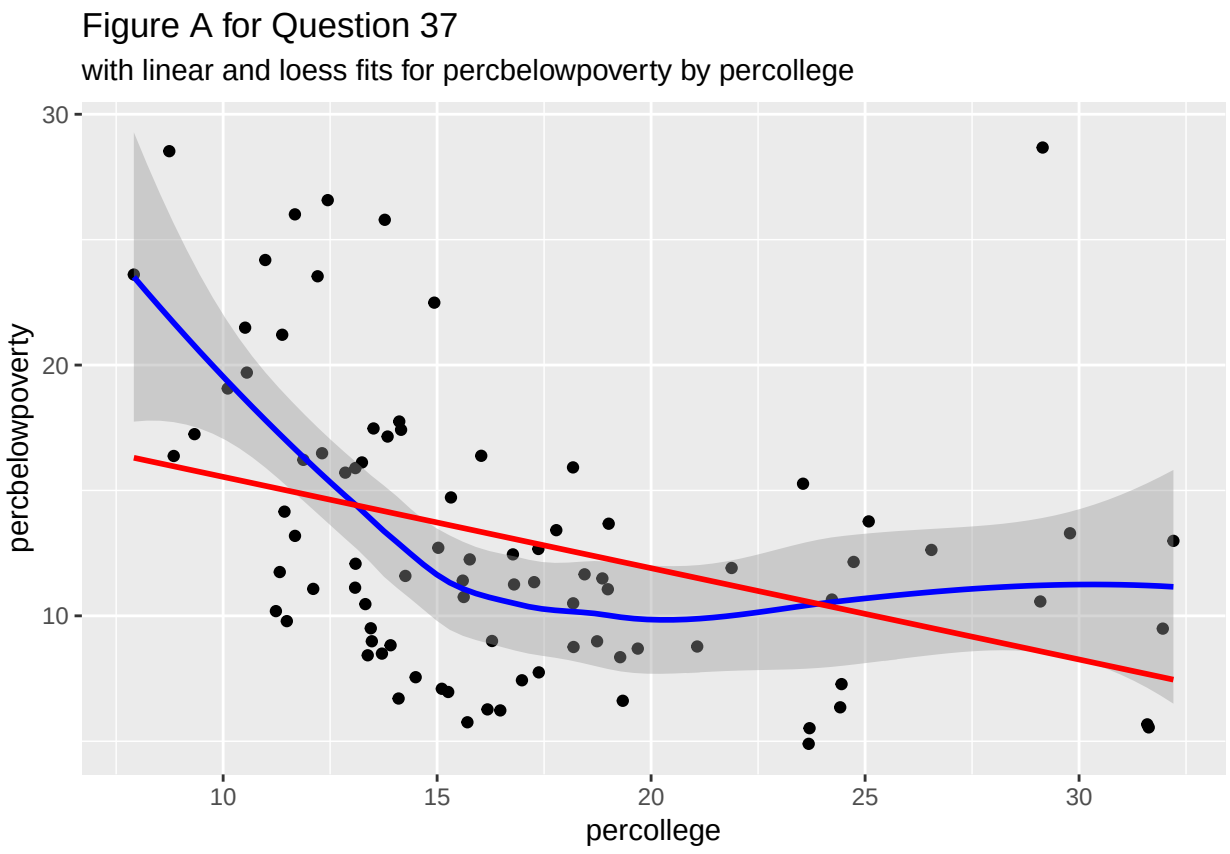


Figure B for Question 37

```
midwest %>% filter(state == "OH") %>%
    ggplot(., aes(x = percollege, y = 1/(percbelowpoverty))) +
    geom_point() +
    geom_smooth(method = "loess", col = "blue") +
    geom_smooth(method = "lm", se = FALSE, col = "red") +
    labs(title = "Figure B for Question 37",
         subtitle = "with linear and loess fits for 1/percbelowpoverty by percollege")
```

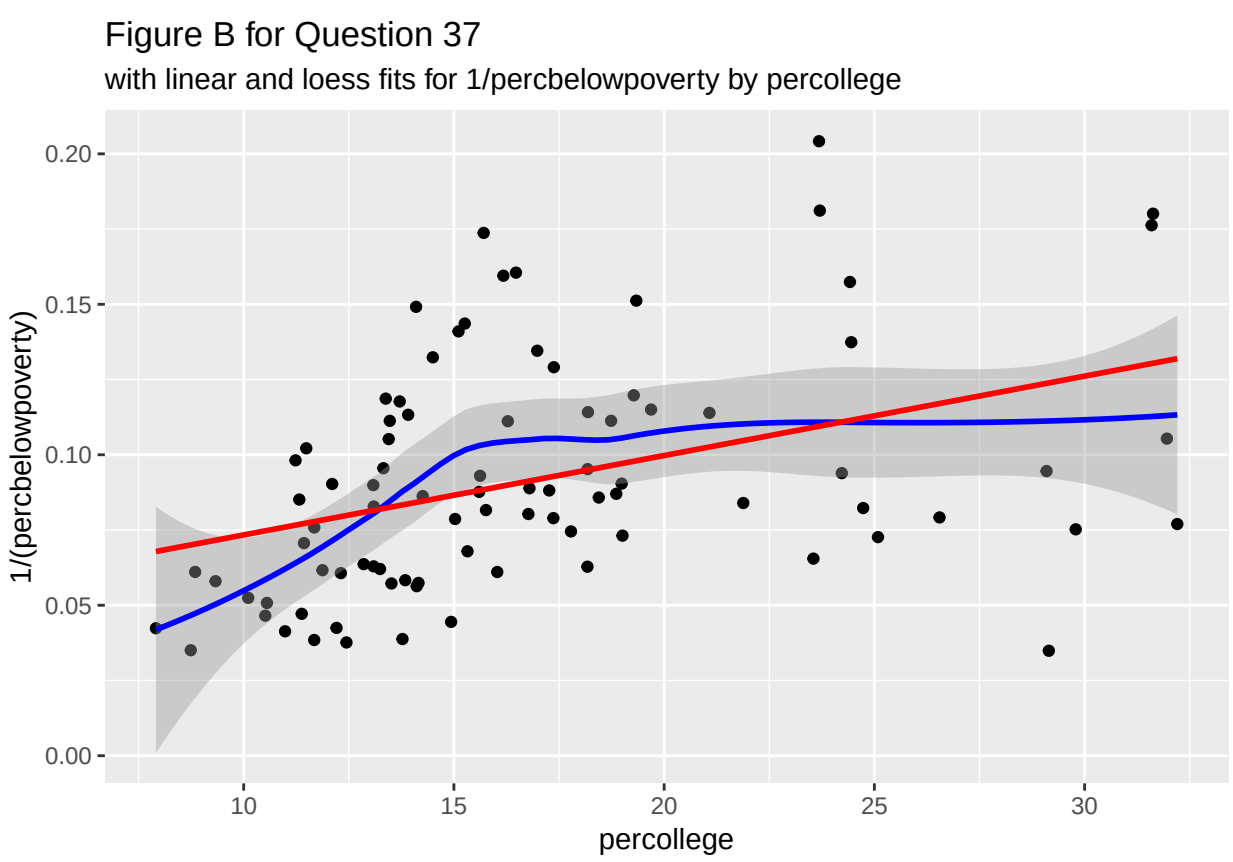


Figure C for Question 37

```
p1 <- midwest %>% filter(state == "OH") %>%
    ggplot(., aes(x = percbelowpoverty)) +
    geom_histogram(bins = 20, fill = "blue", col = "white") +
    labs(title = "Histogram", subtitle = "of percbelowpoverty")

p2 <- midwest %>% filter(state == "OH") %>%
    ggplot(., aes(sample = percbelowpoverty)) +
    geom_qq(col = "blue") + geom_qq_line(col = "purple") +
    labs(title = "Normal Q-Q plot", subtitle = "of percbelowpoverty")

gridExtra::grid.arrange(p1, p2, nrow = 1, top = "Figure C for Question 37")
```

Figure C for Question 37

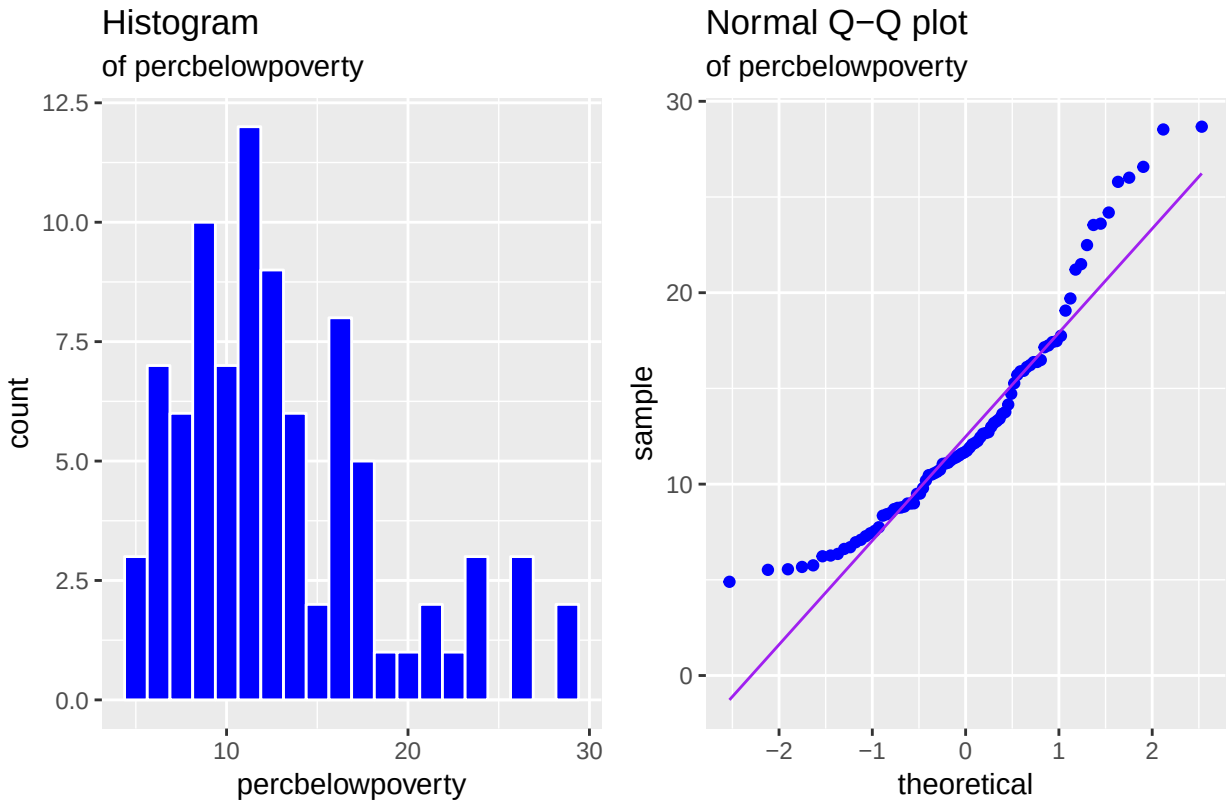
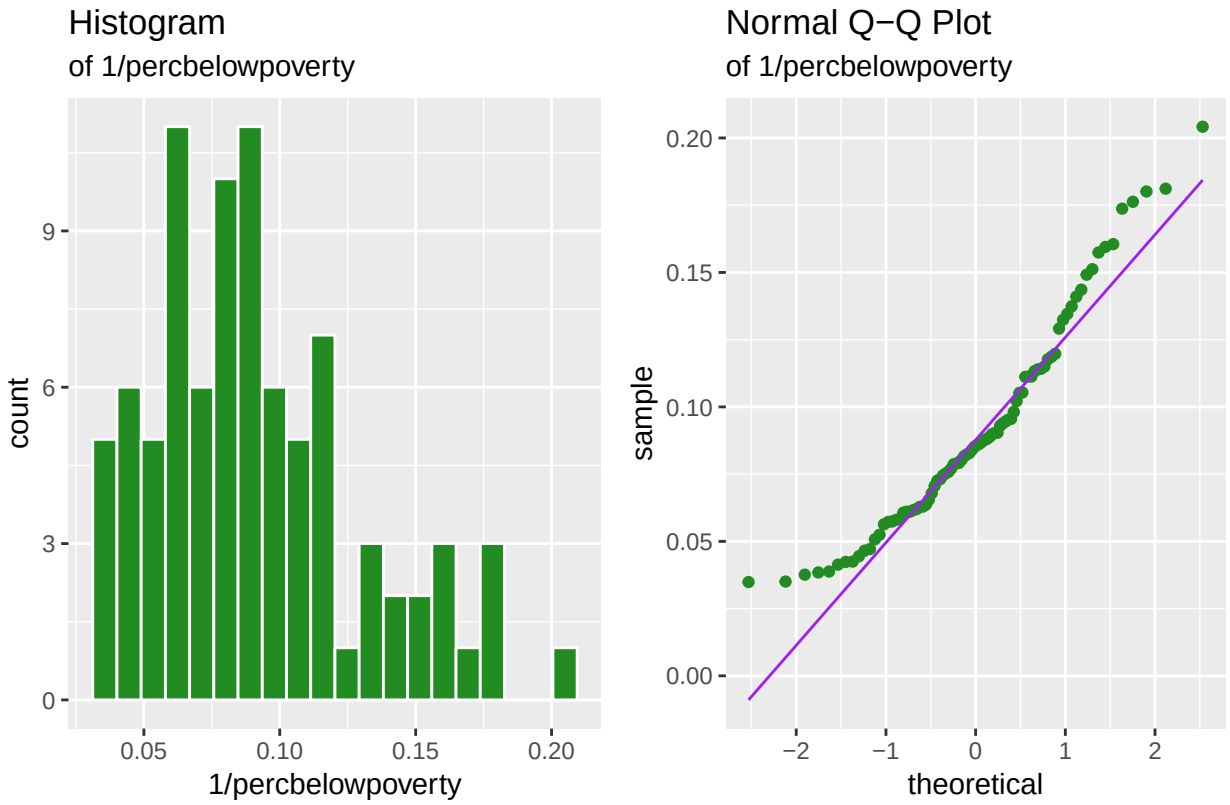


Figure D for Question 37

```
p1 <- midwest %>% filter(state == "OH") %>%
    ggplot(., aes(x = 1/percbelowpoverty)) +
    geom_histogram(bins = 20, fill = "forestgreen", col = "white") +
    labs(title = "Histogram", subtitle = "of 1/percbelowpoverty")

p2 <- midwest %>% filter(state == "OH") %>%
    ggplot(., aes(sample = 1/percbelowpoverty)) +
    geom_qq(col = "forestgreen") + geom_qq_line(col = "purple") +
    labs(title = "Normal Q-Q Plot", subtitle = "of 1/percbelowpoverty")

gridExtra::grid.arrange(p1, p2, nrow = 1, top = "Figure D for Question 37")
```

Figure D for Question 37



Answer for Question 37 is b

Figure B gives us direct evidence as to the impact of choosing an inverse transformation and then fitting a regression model. None of the others do so.

Results on Q37, worth 3 points.

- Only 13/51 (25%) gave the correct response, which was very surprising to me.
- I'm not really sure why option `d` was enticing to so many people. How does Normality play into the question of whether the regression model works?
- There was no partial credit given for this question.

Responses	`b`	`d`	`e`	All Others
Count	13	23	12	3

Question 38

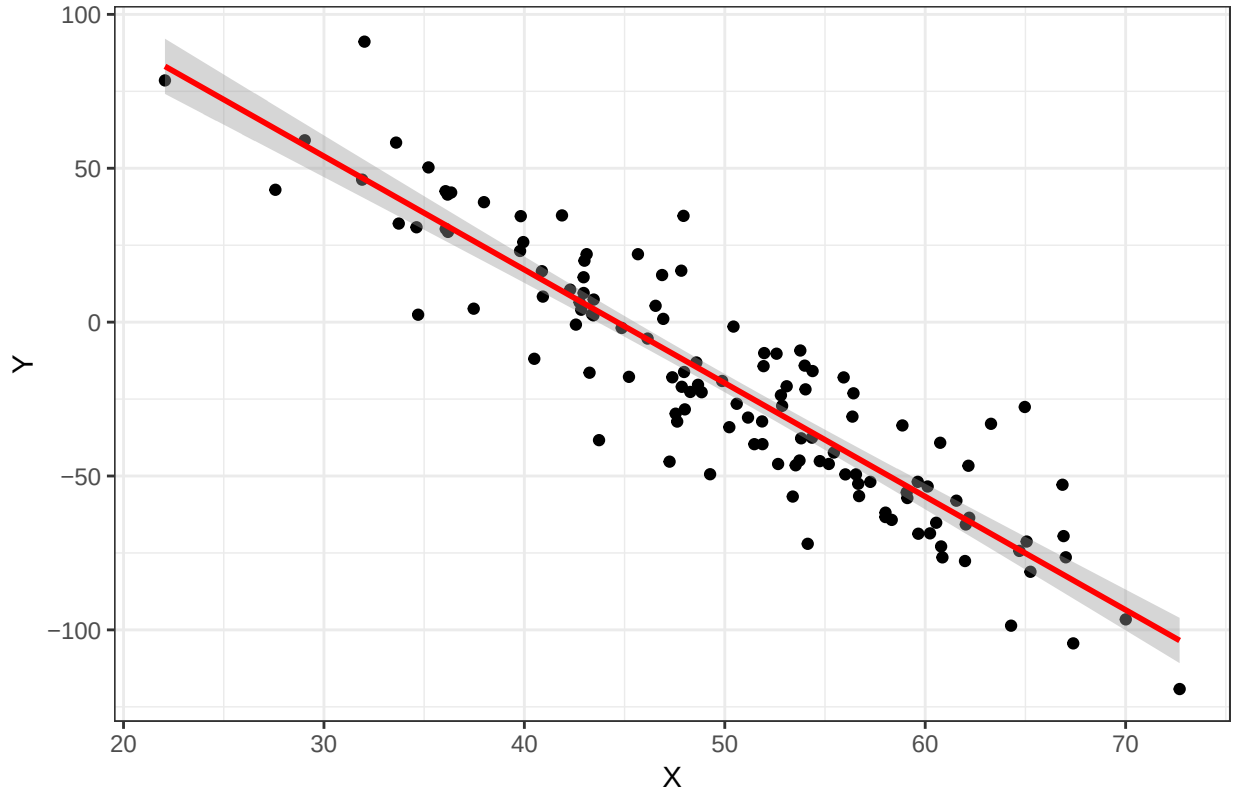
Consider the following possible summaries of a linear model fit to predict Y from X, describing the scatterplot shown in the Figure for Question 38. Which of these summaries is correct?

a. Model: y = 164 - 3.7 x, with R-squared = -0.83
b. Model: y = 164 - 3.7 x, with R-squared = -0.33
c. Model: y = 164 + 3.7 x, with R-squared = 0.83
d. Model: y = 164 + 3.7 x, with R-squared = 0.33
e. Model: y = 164 - 3.7 x, with R-squared = 0.83
f. Model: y = 164 - 3.7 x, with R-squared = 0.33

Figure for Question 38

```
set.seed(43138)
x <- rnorm(125, mean = 50, sd = 10)
err <- rnorm(125, mean = 0, sd = 20)
y = 150 - 3.4*x + err

dat38 <- data_frame(x, y)
ggplot(dat38, aes(x = x, y = y)) +
    geom_point() +
    geom_smooth(method = "lm", col = "red") +
    theme_bw() +
    labs(title = "Figure for Question 38",
         x = "X", y = "Y")
```

Figure for Question 38



Answer for Question 38 is e

- R^2 cannot be negative so a and b are nonsense.
- The Y-X slope is clearly negative (as X increases, Y decreases) so c and d are incorrect
- The cloud of points is tight around the line, and R^2 of 0.83 is far more plausible than 0.33 as a result.

As a demonstration, here is the actual fit.

```
summary(lm(y ~ x, data = dat38))
```

```
Call:
lm(formula = y ~ x, data = dat38)

Residuals:
    Min      1Q  Median      3Q     Max
-41.666 -11.090  -1.768  11.610  47.405

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 164.4579     7.7717   21.16   <2e-16 ***
x            -3.6853     0.1517  -24.30   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.81 on 123 degrees of freedom
```

```
Multiple R-squared:  0.8276,    Adjusted R-squared:  0.8262
F-statistic: 590.3 on 1 and 123 DF,  p-value: < 2.2e-16
```

Results on Q38, worth 3 points.

- 35/51 (69%) gave the correct response.
- The most common incorrect response was `a`, which was disappointing.
- There was no partial credit for this question.

Question 39

Which of the following will create a sample in R of 1000 observations from a Normal distribution with mean of 25 and standard deviation of 4, and place them in a variable called scores. You can assume that the tidyverse package is already loaded, and that an appropriate random seed has been set in a previous command. CHECK ALL THAT APPLY.

a. scores <- 1000*rnorm(n = 1, mean = 25, sd = 4)
b. scores <- rep(rnorm(mean = 25, sd = 4), 1000)
c. scores %>% rnorm(n = 1000, mean = 25, sd = 4)
d. scores <- data_frame(observations = rnorm(1000, mean = 25, sd = 4))
e. scores <- data_frame(rnorm(n = 25, mean = 25, sd = 4))
f. None of these commands.

Answer to Question 39 is f.

None of these commands will do what I asked for. What you need is something like

```
scores <- rnorm(n = 1000, mean = 25, sd = 4)
```

- `a` will produce a single value in `scores` that is a random normal variable multiplied by 1000.
- `b` will produce an error since there's no n value in your rnorm call, among other things.
- `c` uses the pipe %>% rather than the assignment operator <-
- `d` will put the data into a variable called observations within a data frame called scores, and that's not (quite) right.
- `e` produces 25 observations rather than 1000, and puts them in a data frame called scores, rather than a variable called scores.

Results on Q39, worth 2.5 points.

- Only 6/51 respondents (12%) gave the correct response, which was very surprising to me.
- **Partial Credit**: The most common incorrect response was `d` and I decided eventually to give 1 point to everyone who gave that response (and only that response.)

Question 40

1,251 subjects were given a hepatitis C RNA quantitative test which measured the amount of Hepatitis C virus present in their blood, in IU/ml. This measurement is called the viral load, abbreviated load in what follows. Anything over 800,000 is usually considered high, and anything under that is low. Those with low viral load have a better chance of responding to treatment. Consider the two sets of figures for Question 35, below. If our goal is to obtain a transformation of the data which is well fit by a Normal model, which of the following options appears to be our best choice?

a. Taking the square of the viral load.
b. Taking the viral load, untransformed.
c. Taking the natural logarithm of the viral load.
d. Taking the inverse of the viral load.
e. None of these options.

Figure for Question 40

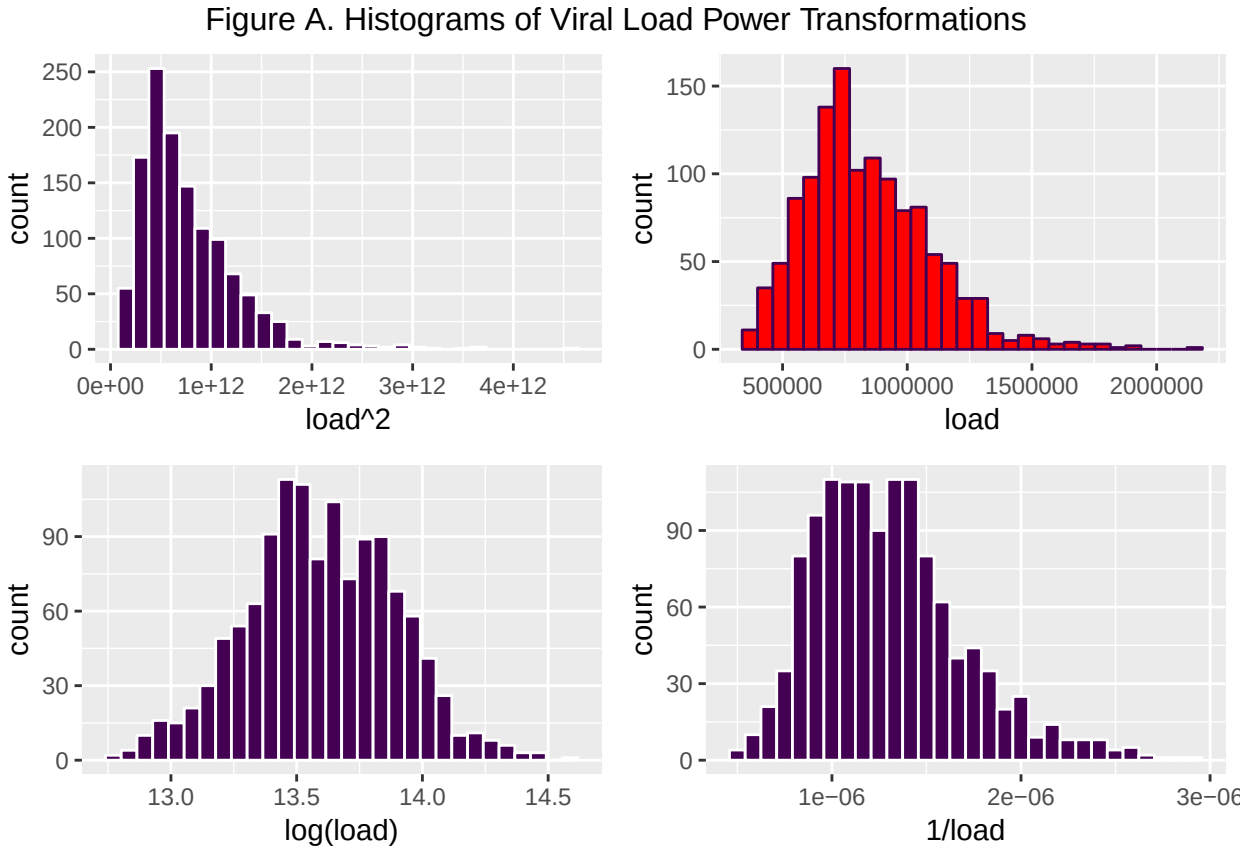
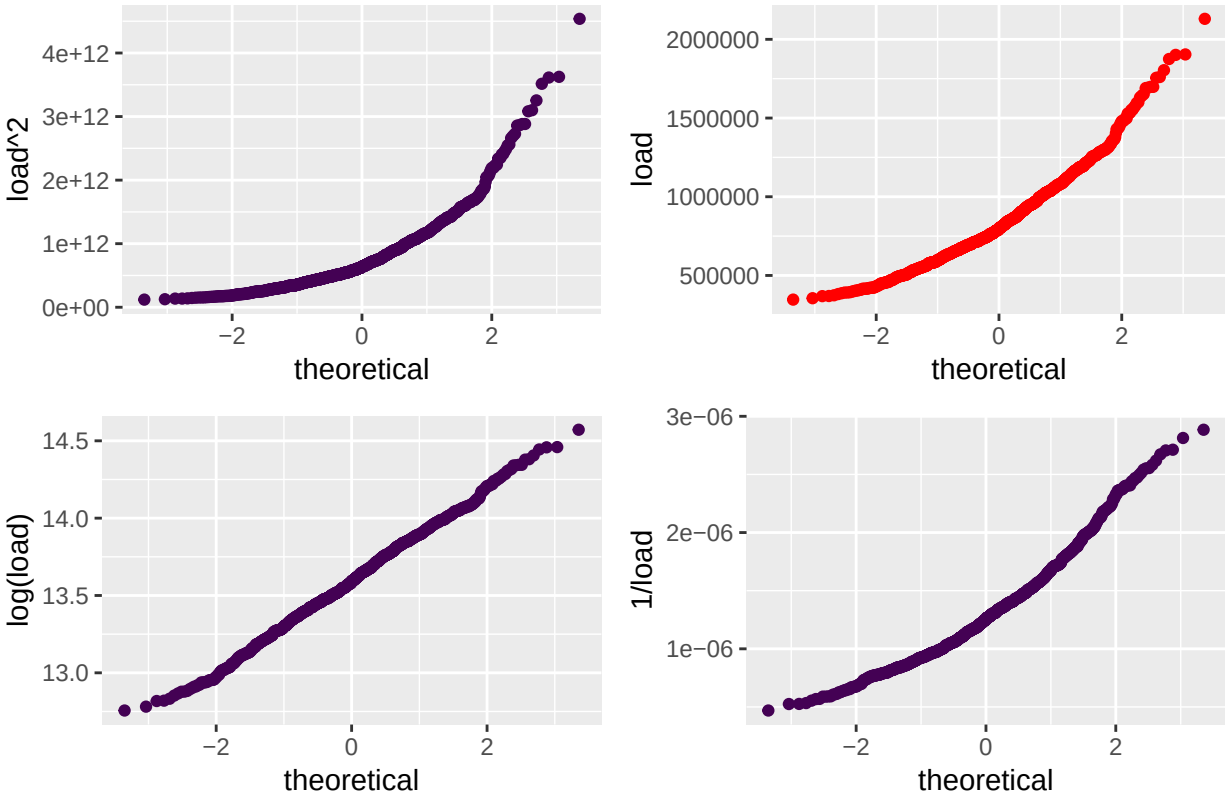


Figure B. Normal Q−Q Plots of Viral Load Power Transformations



Answer for Question 40 is c.

The log transformation is the best choice here. It's the only one that produces a symmetric histogram, or a straight line in the Normal Q-Q plot.

- Note that log produces the natural logarithm (base e) in R. To get the base 10 logarithm, you'd use log10 and to get the base 2 logarithm, you'd use log2. Any logarithm will have the same impact on the model, though.

Results on Q40, worth 2 points.

- At least 46/51 (90% or more) gave the correct response. Good!
- The people who got this wrong selected `d`.
- There was no partial credit for this question.

Answer Key and Results Summary

- `Value` = # of available points on item
- `Item` = Question #
- `Response` = Correct Respopnse
- `Correct` = No. of students (out of 51) who correctly responded (46+ indicates 46-50)
- `% Points` = % of points awarded (including partial credit, which was available on Items Q06, Q10, Q17, Q19, Q21, Q22, Q23, Q24, Q27, Q29, Q30, Q34 and Q39)

Value	Response	Item	Correct	% Points
2 pts	84	Q01	46+	90
3 pts	b	Q02	38	75
2 pts	d	Q03	46+	90
2 pts	d	Q04	46+	94
2.5 pts	c	Q05	44	86
2 pts	c, only	Q06	46+	97
3 pts	d	Q07	43	84
2 pts	15.1	Q08	51	100
2 pts	d	Q09	46+	90
3 pts	-6.8	**Q10**	**36**	**73**
2 pts	d	**Q11**	**21**	**41**
2 pts	a	Q12	50	98
3 pts	b	**Q13**	**36**	**71**
2 pts	d, only	Q14	46+	94
2 pts	d	Q15	46+	90
2.5 pts	c	Q16	39	76
3 pts	a, e	Q17	35	83
2 pts	b	Q18	38	75
3 pts	c, only	Q19	39	84
2 pts	a	Q20	46+	90
2 pts	a, c	**Q21**	**19**	**61**
3 pts	Match x-axis scales	Q22	42	88
3 pts	64.1	**Q23**	**37**	**74**
3 pts	75	Q24	43	86
2.5 pts	b, c	**Q25**	**36**	**71**

Value	Response	Item	Correct	% Points
0.5 pts	Nominal Categorical	Q26a	51	100
0.5 pts	Quantitative	Q26b	51	100
0.5 pts	Quantitative	Q26c	46+	92
0.5 pts	Nominal Categorical	Q26d	51	100
0.5 pts	Ordinal Categorical	Q26e	51	100
3 pts	Add axis labels	Q27	45	90
2 pts	b	**Q28**	**31**	**61**
3 pts	lm(sbp.post ~ sbp.pre, data = dat29)	Q29	34	75
3 pts	facet_wrap(~ nyha, labeller = “label_both”)	**Q30**	**4**	**61**

Value	Response	Item	Correct	% Points
2 pts	a	Q31	46+	94
2.5 pts	d	**Q32**	**19**	**37**
3 pts	12	Q33	44	86
3 pts	a,b,c,d,e	**Q34**	**31**	**71**
3 pts	c	Q35	42	82
2 pts	c	Q36	46+	98
3 pts	b	**Q37**	**13**	**25**
3 pts	e	**Q38**	**35**	**69**
2.5 pts	f	**Q39**	**6**	**35**
2 pts	c	Q40	46+	96

- The 13 bolded items in the tables above had less than 75% of available points awarded.

Your Final Score = Raw Points Total + 8

Add up the points you received on each of the 40 questions, then add 8 additional points to get your final score on the Quiz.

- The highest final scores (by a pair of people) were thus 103/100. Congratulations!
- The median score was 87, and the mean was 85.6
- Here's a summary of the distribution:

Final Score	Count	Approximate Grade
100+	2	A+
90 - 99.5	19	A
85 - 89.5	11	A- or B+
75 - 84	9	B
below 75	10	Dr. Love will be in touch