Answer Sketch: 431 Quiz 1: Fall 2019

Thomas E. Love

due 2019-10-14 at Noon, version 2019-10-15

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 Noon on Monday 2019-10-14, and this is a firm deadline, with no one-hour grace period like we have on Homeworks.

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 **30** questions, and each is worth either 3 or 4 points.

The maximum possible score on the quiz is **100** points. Note that the ten questions worth 4 points are: 2, 5, 6, 10, 12, 14, 20, 22, 28 and 30. They are marked to indicate this.

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

Note that the `dat10.csv` and the `dat22.csv` data files (described in Questions 10 and 22, respectively) are available as part of the Quiz materials on our web site.

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. Assume in all questions that all 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

Consider the `starwars` tibble that is part of the `dplyr` package in the tidyverse. Use those data to answer the following questions:

a. How many of the characters listed in that tibble are a good match for Professor Love, in that they are listed in the tibble as being of the Human `species`, having brown `hair_color` and blue `eye_color`? (Note that we ask for blue `eye_color` and brown `hair_color`, specifically, here, and not other related colors or combinations of these with other colors.)

b. Of the characters you identified in Question 01 part a., how many are from the `homeworld` of Tatooine?

c. How many of the characters in the entire `starwars` tibble have missing data in at least one of the four variables: `species`, `hair_color`, `eye_color` and `homeworld`?

Answer 01a. is 4, 01b. is 2, and 01c. is 18.

For part a., there are four characters who meet these requirements, and for part b., two of them call Tatooine home, as we can see from the output below.

```
starwars %>%
  filter(species == "Human" &
           hair_color == "brown" &
           eye_color == "blue") %>%
  select(name, species, hair_color, eye_color, homeworld)
```

```
# A tibble: 4 x 5
  name               species hair_color eye_color homeworld
  <chr>              <chr>   <chr>      <chr>     <chr>
1 Beru Whitesun lars Human   brown      blue      Tatooine
2 Jek Tono Porkins   Human   brown      blue      Bestine IV
3 Qui-Gon Jinn       Human   brown      blue      <NA>
4 Cliegg Lars        Human   brown      blue      Tatooine
```

For part c., we want something like this:

```
starwars %>%
  count(is.na(species) | is.na(hair_color) | is.na(eye_color) | is.na(homeworld))
```

```
# A tibble: 2 x 2
  `is.na(species) | is.na(hair_color) | is.na(eye_color) | is.na(hom~     n
  <lgl>                                                                <int>
1 FALSE                                                                   69
2 TRUE                                                                    18
```

or perhaps

```
starwars %>%
  filter(!complete.cases(species, hair_color, eye_color, homeworld)) %>%
  nrow()
```

```
[1] 18
```

or maybe

```
starwars %>%
  count(!complete.cases(species, hair_color, eye_color, homeworld))
```

```
# A tibble: 2 x 2
  `!complete.cases(species, hair_color, eye_color, homeworld)`     n
  <lgl>                                                        <int>
1 FALSE                                                           69
2 TRUE                                                            18
```

which gives the answer directly, but there are many, many ways to get it a little less directly, such as by first using

```
nrow(starwars)
```

```
[1] 87
```

to tell you that there are 87 characters in the tibble as a whole, and then something like:

```
starwars %>%
  filter(complete.cases(species, hair_color, eye_color, homeworld)) %>%
  nrow()
```

```
[1] 69
```

to tell you that 69 of them have complete data on these four variables, and thus the other 18 do not.

Results

Item	01a	01b	01c
Correct (out of 60)	> 54	> 54	47
% of Available Points Awarded	> 90	> 90	78

Each part (a, b and c) was worth 1 point, without partial credit.

I asked for numerical responses. Some instead wrote sentences (or things like Four, Two and Eighteen.) Others wrote out the names of the characters, sometimes without specifying the counts. All of that was unnecessary, but didn't cost you any points.

Question 02 (4 points)

Suppose that you built a subset of the `starwars` data called `humanbrown` which consists only of the characters who are Human with brown `eye_color`. Now, you 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(humanbrown %>% select(mass))
b. humanbrown %>% filter(complete.cases(mass)) %>% summarize(quantile(mass, probs = 0.5))
c. humanbrown %>% summarize(median(mass, na.rm = TRUE))
d. humanbrown %>% filter(complete.cases(mass)) %>% summarize(median(mass))
e. mosaic::favstats(~ mass, data = humanbrown)
f. None of these.

Answer 02 is that all of them (a, b, c, d, and e) work.

First, we'll build the subset.

```
humanbrown <- starwars %>% filter(species == "Human", eye_color == "brown")
```

Then we'll try it. The answer turns out to be 79, and all five approaches work.

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

```
      mass
 Min.   :45.00
 1st Qu.:78.60
 Median :79.00
 Mean   :74.75
 3rd Qu.:82.00
 Max.   :85.00
 NA's   :6
```

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

```
# A tibble: 1 x 1
  `quantile(mass, probs = 0.5)`
                          <dbl>
1                            79
```

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

```
# A tibble: 1 x 1
  `median(mass, na.rm = TRUE)`
                         <dbl>
1                           79
```

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

```
# A tibble: 1 x 1
  `median(mass)`
           <dbl>
1             79
```

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

```
 min   Q1 median Q3 max     mean       sd  n missing
  45 78.6     79 82  85 74.74545 13.94822 11       6
```

Results

Item	02
Correct (out of 60)	48
% of Available Points Awarded	88

Partial credit was awarded according to the following schedule.

Response	Points Awarded
a, b, c, d, e	4
a, b, c, d	3
a, c, d, e	3
a, c, e	2
a, e	1
b, c, d	2
b, c, d, e	3
b, d	1
c	0
d	0
f	0

Question 03

I produced the cross-tabulation shown in the Output for Question 03 using the complete `starwars` tibble available in the `tidyverse`. All relevant packages are loaded on my computer. Which one of the following commands did I use?

a. mosaic::favstats(~ gender + height < 160, data = starwars)
b. starwars %$% table(gender, height < 160)
c. starwars %$% tabyl(gender, height < 160)
d. starwars %$% filter(height < 160) %>% count(gender)
e. table(gender, height < 160, data = starwars)
f. None of these would work.

Output for Question 03

```
gender          FALSE TRUE
  female           13    4
  hermaphrodite     1    0
  male             52    7
  none              1    0
```

Answer 03 is b

I obtained this result with:

```
starwars %$% table(gender, height < 160)
```

```
gender          FALSE TRUE
  female           13    4
  hermaphrodite     1    0
  male             52    7
  none              1    0
```

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

Results

Item	03
Correct (out of 60)	> 54
% of Available Points Awarded	> 90

There was no partial credit available on this question.

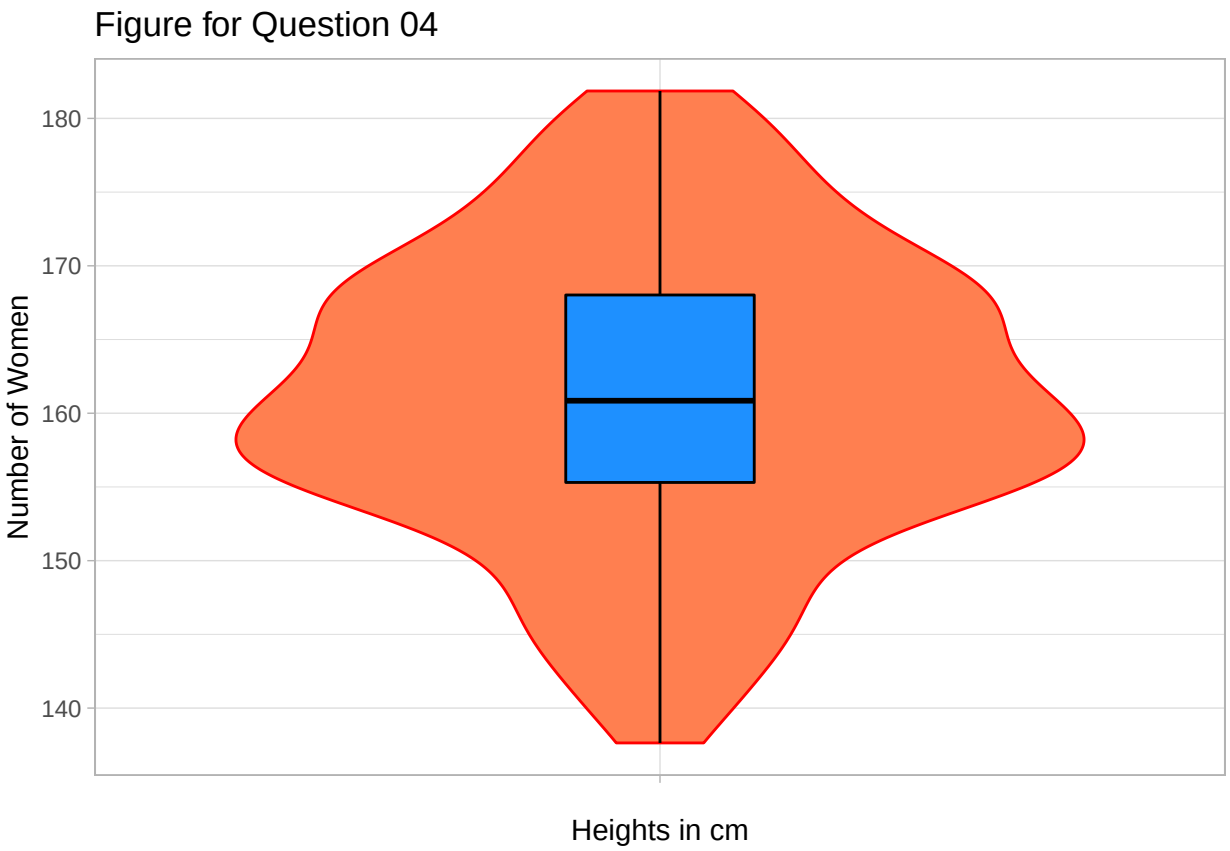
Question 04

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 `dat04`. Which of the following bits of R code was NOT used in generating the Figure for Question 04?

The response options for Question 04 are:

a. geom_boxplot(fill = "dodgerblue", col = "black", width = 0.2)
b. geom_line(fill = "blue", col = "white")
c. geom_violin(col = "red", fill = "coral")
d. ggplot(data = dat04, aes(x = "", y = height))
e. labs(title = "Figure for Question 04")
f. labs(x = "Heights in cm", y = "Number of Women")
g. theme_light()

Figure for Question 04



Answer 04 is b

The Figure in Question 04 was made by combining the other six bits of code. This is a boxplot and a violin plot, but there's no line chart (as would be implied by `geom_line`) here.

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

```
set.seed(2018004)
temp <- rnorm(148, mean = 161.4, sd = 9.2)

dat04 <- tibble(height = temp)

ggplot(data = dat04, aes(x = "", y = height)) +
    geom_violin(col = "red", fill = "coral") +
    geom_boxplot(fill = "dodgerblue", col = "black", width = 0.2) +
    labs(x = "Heights in cm", y = "Number of Women") +
    labs(title = "Figure for Question 04") +
    theme_light()
```

Results

Item	04
Correct (out of 60)	> 54
% of Available Points Awarded	> 90

There was no partial credit available on this question.

Question 05 (4 points)

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 shown in the Figure for Question 05 does not include a problem of this sort. Please identify the **good** plot - the one that avoids Jeff's pitfalls.

a. Plot A
b. Plot B
c. Plot C
d. Plot D

Figure for Question 05

Answer 05 is d

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

- Plot A 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 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 uses a figure title that specifies the type of plot used, without describing the result.
- Plot D is essentially reasonable. It is the best of these four plots.

Results

Item	05
Correct (out of 60)	> 54
% of Available Points Awarded	> 90

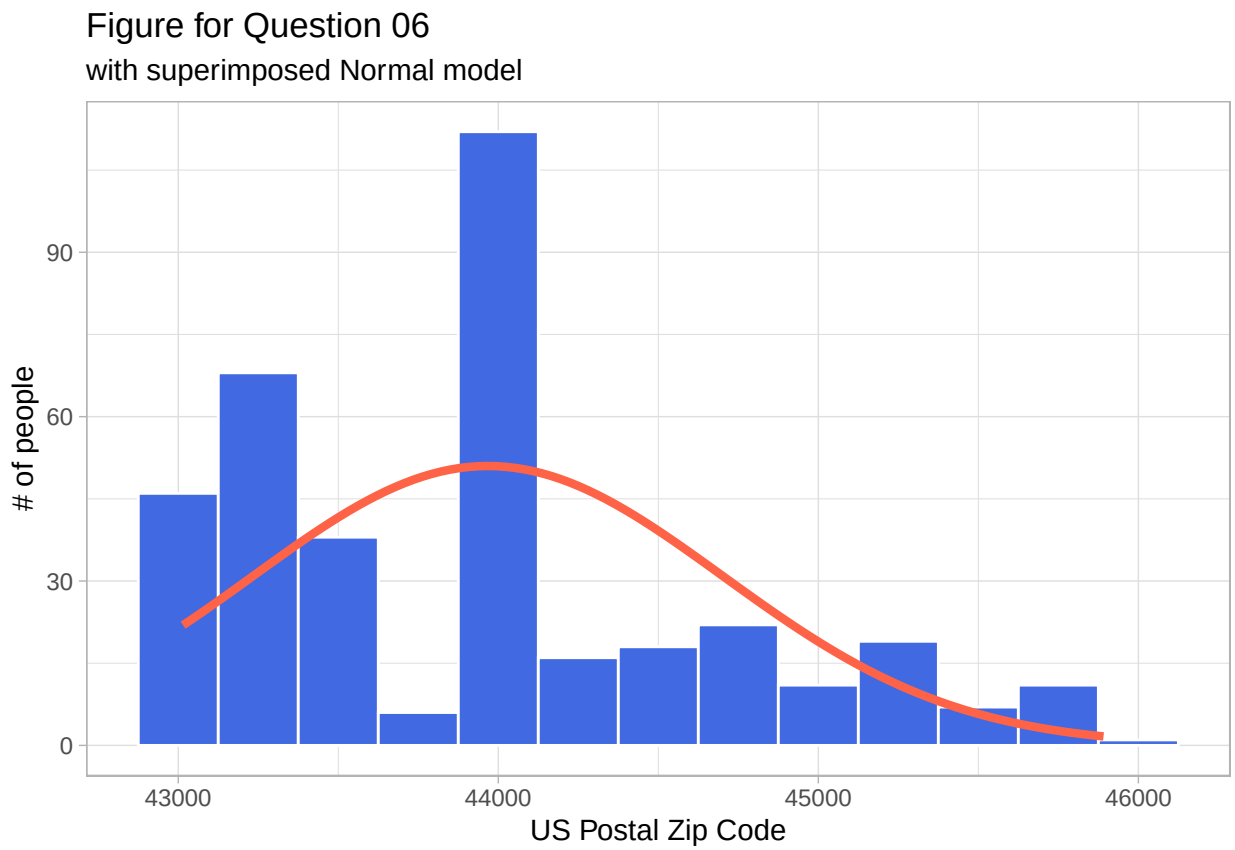
There was no partial credit available on this question.

Question 06 (4 points)

The Figure for Question 06 shows the US postal zip codes of the last 375 people from the state of Ohio to visit a web site providing information on purchasing insurance through the federal Health Insurance Marketplace. Which one of the following summaries of these data would be most appropriate?

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

Figure for Question 06



Answer 06 is c

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

Results

Item	06
Correct (out of 60)	34

Item	06
% of Available Points Awarded	57

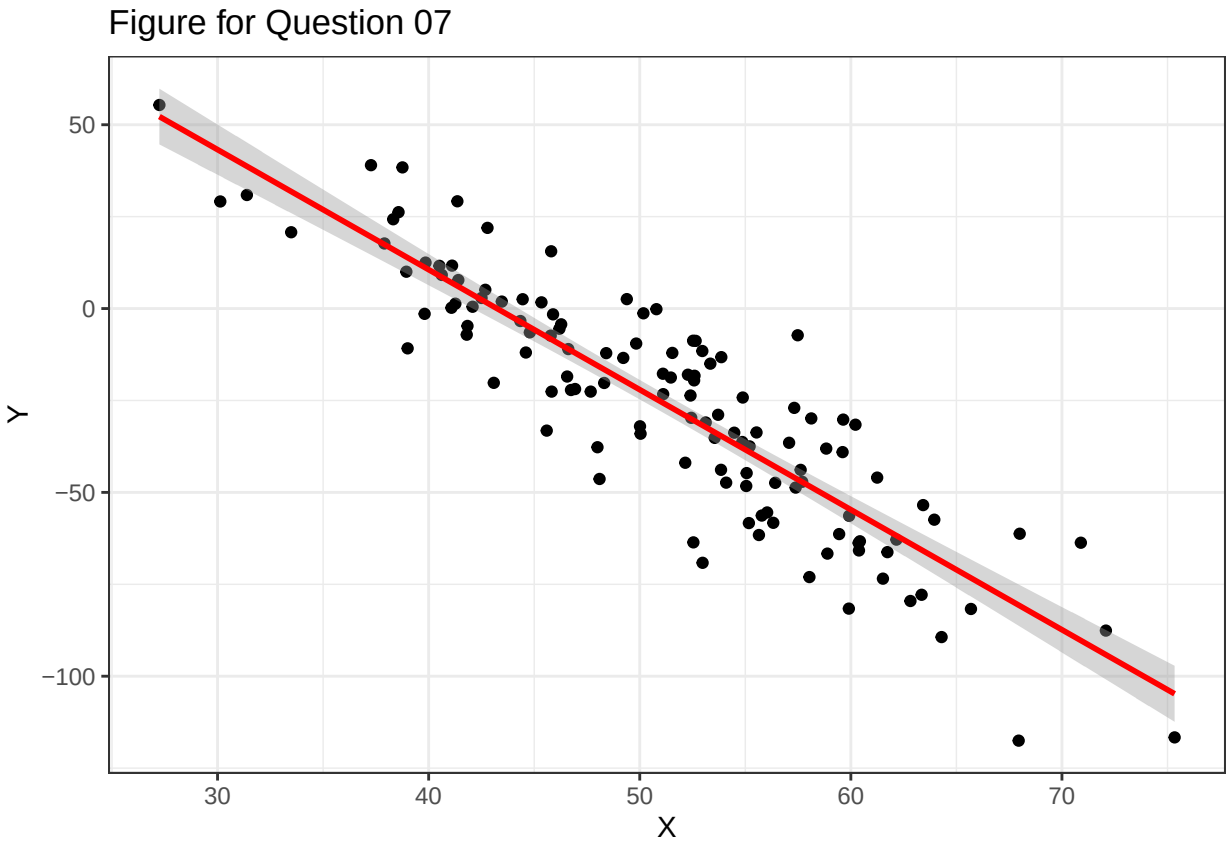
There was no partial credit available on this question. It wasn't a surprise that people found this tricky, but statistics is about the details, sometimes.

Question 07

Consider the following possible summaries of a linear model fit to predict Y from X, describing the scatterplot shown in the Figure for Question 07. Which of these summaries is correct?

a. Model: y = 141 - 3.3 x, with R-squared = -0.79
b. Model: y = 141 - 3.3 x, with R-squared = -0.29
c. Model: y = 141 + 3.3 x, with R-squared = 0.79
d. Model: y = 141 + 3.3 x, with R-squared = 0.29
e. Model: y = 141 - 3.3 x, with R-squared = 0.79
f. Model: y = 141 - 3.3 x, with R-squared = 0.29
g. Model: y = 3.3 + 141 x, with R-squared = -0.79
h. Model: y = 3.3 - 141 x, with R-squared = -0.29
i. Model: y = -3.3 + 141 x, with R-squared = 0.79
j. Model: y = -3.3 + 141 x, with R-squared = 0.29
k. Model: y = 3.3 + 141 x, with R-squared = 0.79
l. Model: y = 3.3 + 141 x, with R-squared = 0.29

Figure for Question 07



Answer 07 is e

- R^2 cannot be negative so a and b and k and l are nonsense.
- The Y-X slope is clearly negative (as X increases, Y decreases) so c and d are incorrect

- The models proposed in g, h, i, j, k and l all get the slope and intercept backwards
- The cloud of points is tight around the line, and R^2 of 0.79 is far more plausible than 0.29 as a result.

As a demonstration, here is the actual fit.

```
summary(lm(y ~ x, data = dat07))
```

```
Call:
lm(formula = y ~ x, data = dat07)

Residuals:
    Min      1Q  Median      3Q     Max
-37.368  -9.831   0.471  10.189  39.232

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 141.1720     7.7865   18.13   <2e-16 ***
x            -3.2647     0.1498  -21.79   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.85 on 123 degrees of freedom
Multiple R-squared:  0.7943,    Adjusted R-squared:  0.7926
F-statistic:   475 on 1 and 123 DF,  p-value: < 2.2e-16
```

Results

Item	07
Correct (out of 60)	43
% of Available Points Awarded	72

There was no partial credit available on this question.

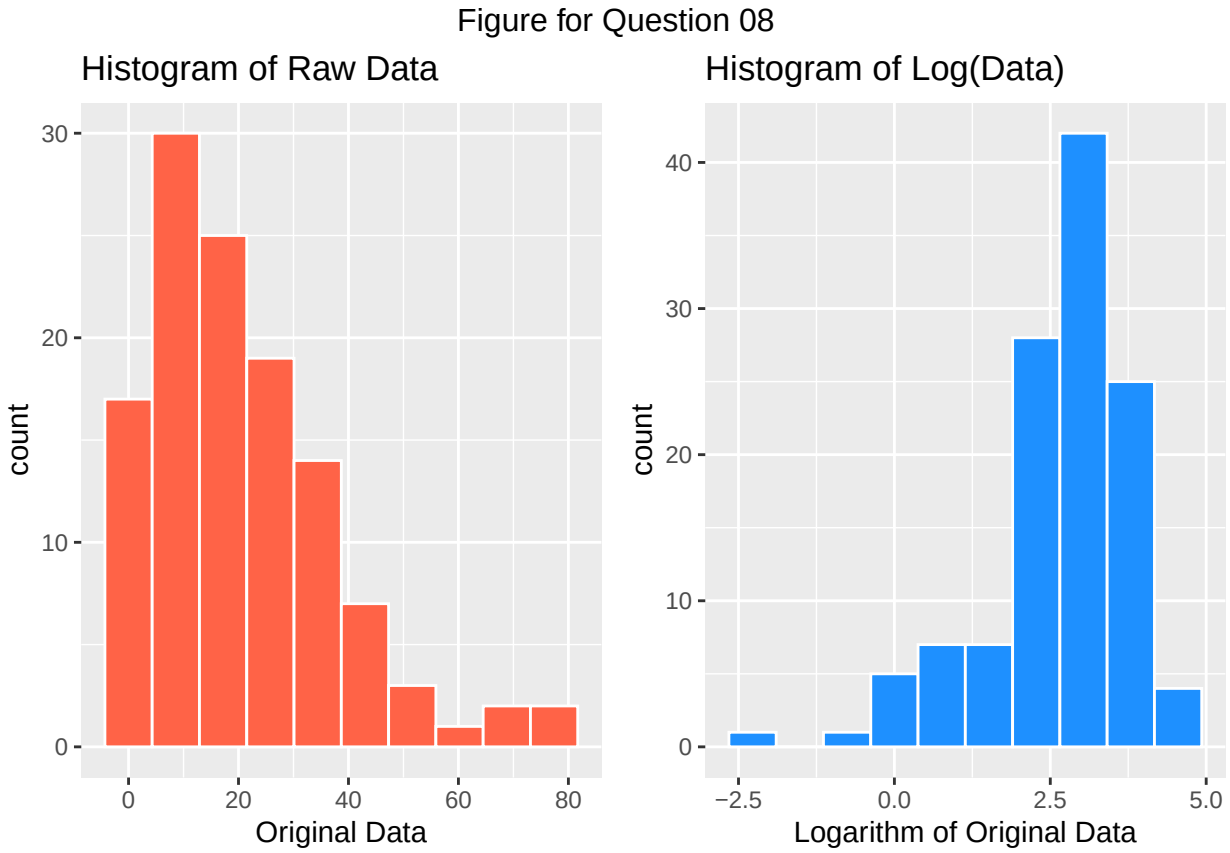
Question 08

Consider the two histograms shown in the Figure for Question 08. 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 08 are:

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

Figure for Question 08



Answer 08 is a

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

Item	08
Correct (out of 60)	49
% of Available Points Awarded	82

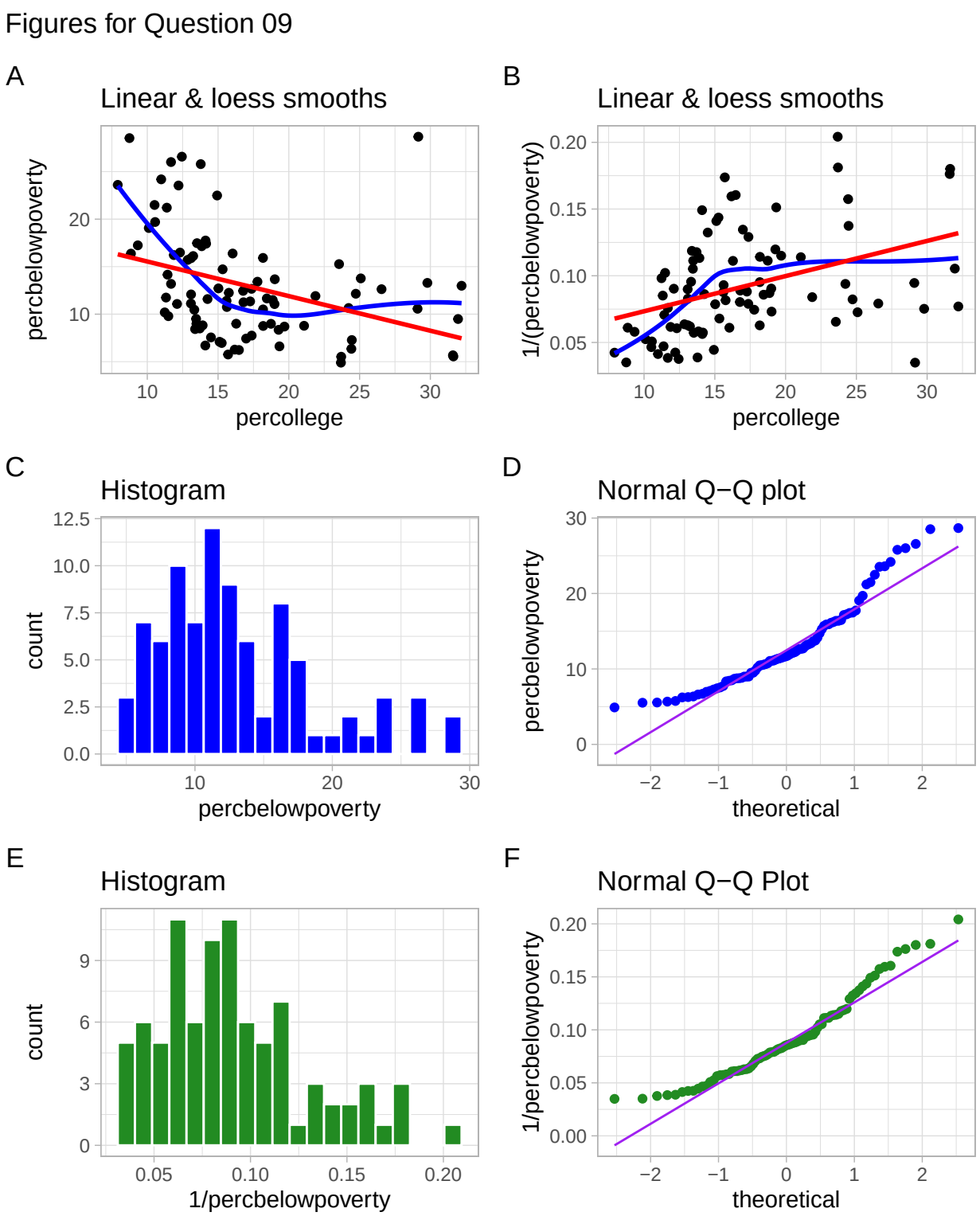
There was no partial credit available on this question.

Question 09

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 Plots shown in the Figures for Question 09 would be of the most help in assessing whether using this transformation would improve the assumption of linearity?

a. Plot A
b. Plot B
c. Plots C and D
d. Plots E and F
e. They would all be equally useful

Figures for Question 09



Answer 09 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

Item	09
Correct (out of 60)	9
% of Available Points Awarded	15

This went quite poorly, as you can see. The most common incorrect response was the combination of Plots E and F, which certainly address the assumption of Normality in the residuals, but tell us nothing about linearity.

There was no partial credit available on this question.

Setup for Questions 10-12

Questions 10-12 make use of the `dat10` data that describe 40 patients with either aortic or mitral regurgitation who had heart surgery.

The data are stored in the `dat10` tibble, as shown. The variables are:

- `subj_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.

Note that the `dat10.csv` data file is available to you as part of the Quiz materials.

Tibble (with Code) for Questions 10-12

```
dat10 <- read.csv("data/dat10.csv") %>% tbl_df

dat10
```

```
# A tibble: 40 x 7
   subj_id ef_pre ef_post reg_type NYHA  sbp_pre sbp_post
   <fct>    <dbl>   <dbl> <fct>    <fct>   <int>    <int>
 1 S-127     0.56    0.34 aortic   III       140      138
 2 S-156     0.62    0.47 aortic   II        110      110
 3 S-174     0.67    0.41 mitral   II        120      120
 4 S-175     0.63    0.48 mitral   III        90      100
 5 S-222     0.74    0.54 mitral   II        130      115
 6 S-263     0.6     0.33 aortic   III       150      135
 7 S-288     0.53    0.47 aortic   I         215      130
 8 S-298     0.69    0.6  mitral   II         95      100
 9 S-300     0.6     0.3  aortic   III       150      130
10 S-341     0.66    0.43 mitral   IV        125      124
# ... with 30 more rows
```

Question 10 (4 points)

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 dat10 tibble. Be sure that your code will work, and in particular, that you haven't spelled anything incorrectly.

Answer 10 is a line of R code, like `lm(sbp_post ~ sbp_pre, data = dat10)`

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 = dat10)`
- `coef(lm(sbp_post ~ sbp_pre, data = dat10))`
- `summary(lm(sbp_post ~ sbp_pre, data = dat10))`
- `broom::tidy(lm(sbp_pre ~ sbp_post, data = dat10))`

Results

Item	10
Correct (out of 60)	42
% of Available Points Awarded	74

Examples of incorrect responses include:

- placing the result of the model in an object, but not printing it, like `model <- lm(sbp_post ~ sbp_pre, data = dat10)` (you got 2 points of partial credit if you did this)
- giving me the result, in addition to partially correct code (I gave 2 points of partial credit for this)
- attempts to summarize the correlation, rather than the regression model, like `cor(dat10$sbp_post, dat10$sbp_pre)` wouldn't help.
- fitting the model backwards, as in `lm(data10$sbp_pre ~ data10$sbp_post)` which wouldn't help.
- using variable names not in the data set, like `lm(sbp.post ~ sbp.pre, data = dat10)` with a dot rather than an underscore, which wouldn't help.
- looking at the ejection fraction variables rather than the systolic blood pressures wouldn't help, either.

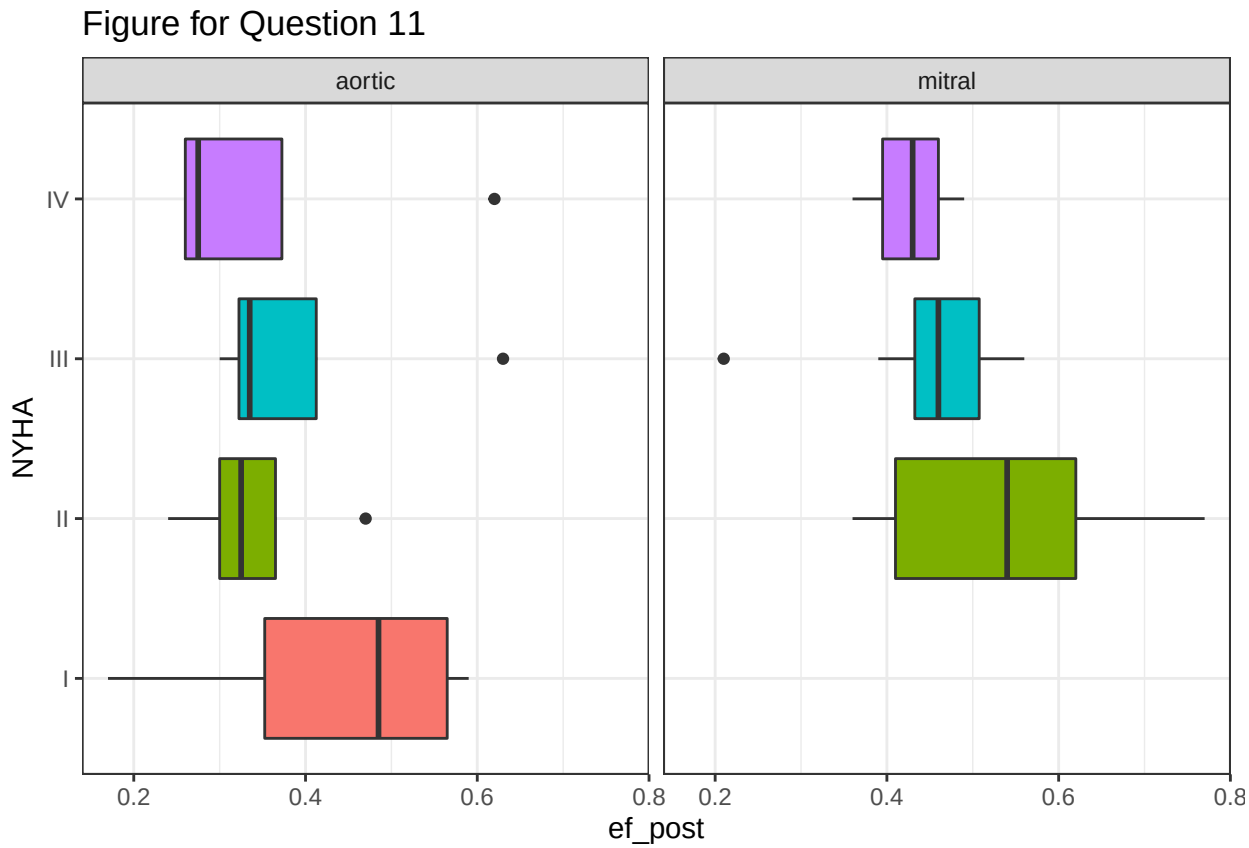
Question 11

The Figure for Question 11 shows the ejection fraction after surgery for 40 patients, and the complete code used to develop the Figure, based on the `dat10` tibble, 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 11 rather than eight?

a. drop_na
b. facet_grid(~ reg_type, labeller = "label_both")
c. summary(dat10)
d. dat10 %>% group_by(NYHA) %>% summarize(reg_type)
e. dat10 %>% count(reg_type, NYHA)
f. None of these would be useful.

Figure (with Code) for Question 11

```
ggplot(dat10, aes(x = NYHA, y = ef_post, fill = NYHA)) +
    geom_boxplot() +
    facet_wrap(~ reg_type) +
    coord_flip() +
    guides(fill = FALSE) +
    labs(title = "Figure for Question 11") +
    theme_bw()
```



Answer 11 is e

As we can see from the results of applying the code in **e** 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 11. None of the other codes would provide us with this information, although there are other ways we could have used to figure this out.

```
dat10 %>% 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

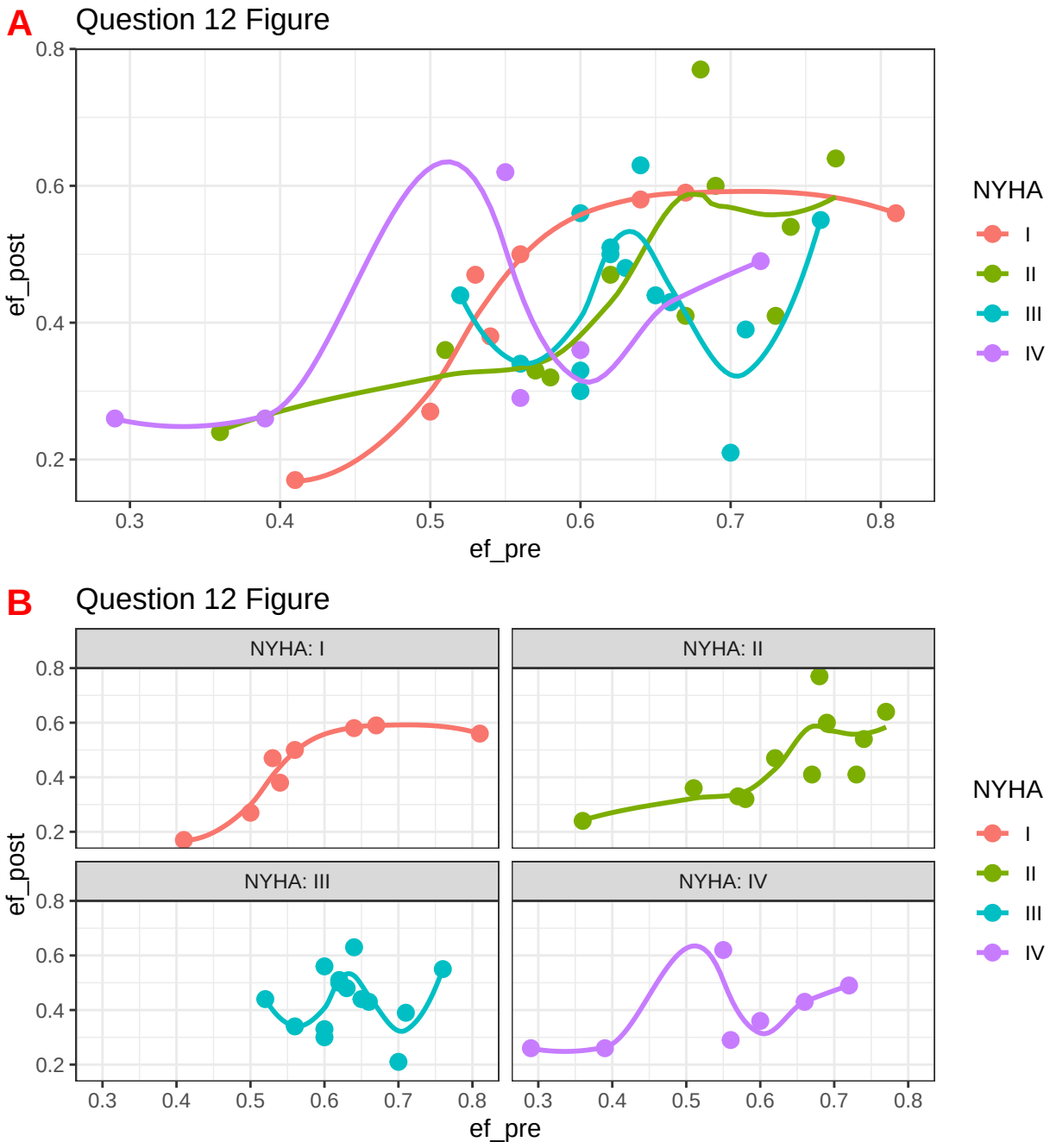
Item	11
Correct (out of 60)	53
% of Available Points Awarded	88

There was no partial credit available on this question.

Question 12 (4 points)

The Figures for Question 12 (labeled A and B) each show the same data, and the same loess smooths, using two different approaches. Please specify the one-line R command used to turn Plot A into Plot B.

Figures for Question 12



Answer 12 is `facet_wrap(~ NYHA, labeller = "label_both")`

The actual code was:

```
p_A <- ggplot(dat10, aes(x = ef_pre, y = ef_post, col = NYHA)) +
  geom_point(size = 3) +
  geom_smooth(method = "loess", se = FALSE) +
  labs(title = "Question 12 Figure") +
  theme_bw()

p_B <- ggplot(dat10, aes(x = ef_pre, y = ef_post, col = NYHA)) +
  geom_point(size = 3) +
  geom_smooth(method = "loess", se = FALSE) +
  facet_wrap(~ NYHA, labeller = "label_both") +
  labs(title = "Question 12 Figure") +
  theme_bw()

plot_grid(p_A, p_B,
          labels = c('A', 'B'),
          label_colour = "red",
          label_size = 16, ncol = 1)
```

Results

Item	12
Correct (out of 60)	12
% of Available Points Awarded	56

Substantial partial credit was available.

- A response of `facet_wrap(~ NYHA)` without the labeller piece got you two points of partial credit, since you would have produced a plot that works, just without the labels specified as desired.
- If you misspecified how labeller works, perhaps with `facet_wrap(~ NYHA, labelled= "label_values")` that got you one point of partial credit, since you're actually committing an error, rather than just not labeling the values.
- If you left out the quotation marks in labeller, like `facet_wrap(~ NYHA, labeller = label_both)`, that, too, would have led to an error, so you got one point of partial credit.
- You didn't need to specify the `ncol` or `nrow` in your `facet_wrap()` command, but if you did so incorrectly, like 'facet_wrap(~ NYHA, ncol = 4, labeller = "label_both"), you lost a point for that detail.
- If you used `nyha` instead of `NYHA` you lost a point for that detail.
- We didn't worry about whether or not you included the + sign before or after the command.

Question 13

In the Tibble for Question 13, are the arm and nose lengths (in cm) and the arm-nose ratio (ANratio), calculated by dividing the arm length by the nose length, for 20 subjects. I have not provided you with an electronic copy of the data set for this question, but you will see a small piece of summary output.

The Statue of Liberty's nose measures 4 feet, 6 inches, and her arm is 42 feet long. Calculate the Statue's 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 subjects listed in the Tibble for Question 13. Your response should be the Z score for the Statue of Liberty, with an appropriate sign, rounded to one decimal place.

Tibble and Output for Question 13

```
dat13
```

```
# A tibble: 20 x 4
   subject   arm  nose ANratio
   <chr>   <dbl> <dbl>   <dbl>
 1 Akari    73.3   4.7    15.6
 2 Beth     60.3   4.5    13.4
 3 Carol    74.7   4.7    15.9
 4 Donna    60.6   4.3    14.1
 5 Early    65.6   4.4    14.9
 6 Feng     63.2   4.3    14.7
 7 Grace    75.7   4.7    16.1
 8 Hanna    67.9   4.3    15.8
 9 Ione     63.4   4.2    15.1
10 Julie    74.4   4.8    15.5
11 Karen    62.6   4.2    14.9
12 Lin      64.1   4.3    14.9
13 Mary     69.7   4.2    16.6
14 Nancy    69.6   4.7    14.8
15 Olive    75.4   4.6    16.4
16 Paris    81.4   5.5    14.8
17 Ruo      70.8   4.4    16.1
18 Sara     82.2   5.2    15.8
19 Tilly    58.5   4.4    13.3
20 Vivi     65.8   4.3    15.3
```

```
dat13 %>% summarize(sum(ANratio))
```

```
# A tibble: 1 x 1
  `sum(ANratio)`
           <dbl>
1            304
```

```
dat13 %$% var(ANratio) %>% round(., 2)
```

```
[1] 0.81
```

Answer 13 is -6.5

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 $42/(4.5) = 9.33$.

As for the 20 women in the tibble, we can obtain their mean and standard deviation:

```
mosaic::favstats(~ ANratio, data = dat13)
```

```
  min   Q1 median     Q3  max mean        sd  n missing
 13.3 14.8   15.2 15.825 16.6 15.2 0.9002924 20       0
```

The mean of the 20 women is 15.2, since the sum we provided was 304, and since 304/20 is 15.2, and the standard deviation is 0.9 (or, if you like, 0.9003 or 0.9002924, but it won't matter.) Of course, we know the standard deviation was about 0.9 because that's the square root of the variance (0.81) that I provided to you. So you didn't actually need to type in any of the data to get what you needed.

Thus, the Z score for the statue is

$$(9.33 - 15.2)/0.9 = -6.522$$

And so Z = -6.5 for the Statue of Liberty after rounding to one decimal place. Note that if we'd used either 0.9003 for the standard deviation, or even 0.9002924, our Z score would still round to -6.5.

Item	13
Correct (out of 60)	31
% of Available Points Awarded	57

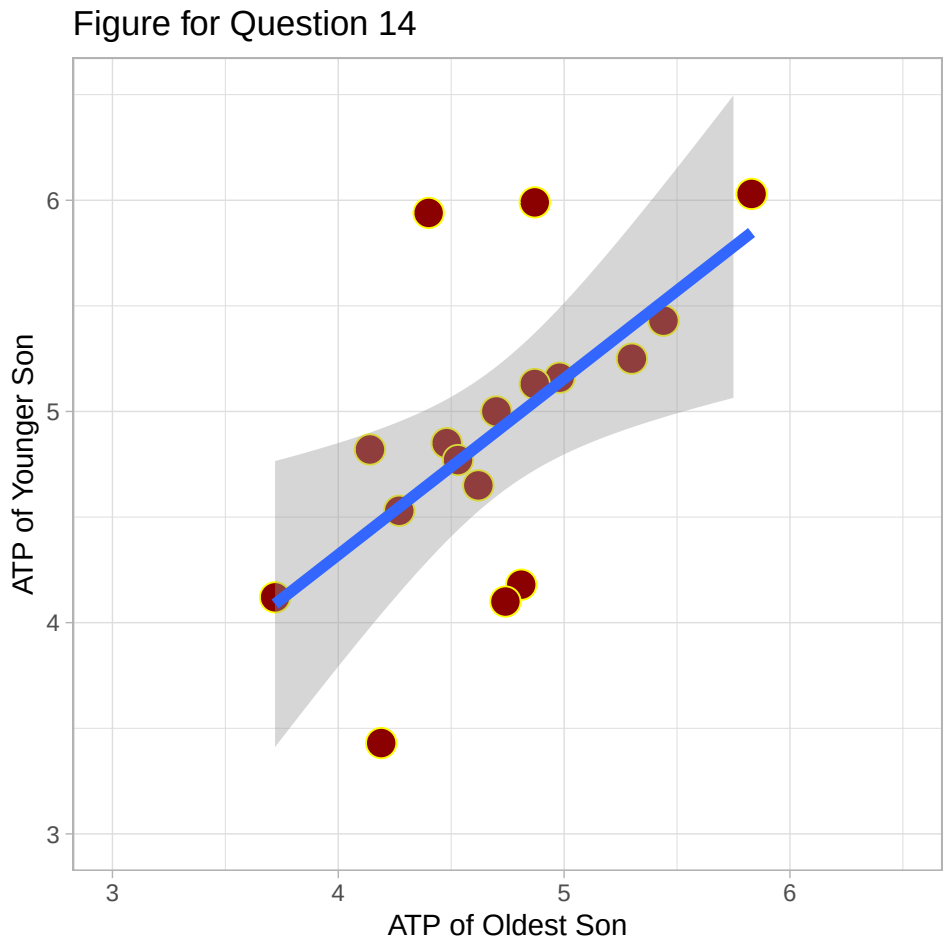
- Since we asked you to round to one decimal place, you got 2 points of partial credit for an answer that was correct, but improperly rounded (say, -6.52).
- Perhaps overly generously, we gave one point of partial credit for -6.6 and -6.4 as responses.
- We just wanted the number. Some people wrote a sentence, or something like "-6.5 standard deviations" or wrote out the details of their calculation. We didn't penalize for that.

Question 14 (4 points)

Dern and Wiorkowski (1469) 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 14 depicts the data for 17 pairs of brothers. Suppose we are also interested in an 18th family, where Kevin is the oldest son and Brian is the youngest son. Which of the following statements are true? (CHECK ALL THAT APPLY.)

a. If Kevin's ATP is 5, the linear model's point estimate for Brian's ATP exceeds 5.
b. The absolute value of the Pearson correlation is between 0 and 0.25.
c. The intercept of the regression line is less than zero.
d. The slope of the regression line is greater than zero.
e. None of these statements are true.

Figure for Question 14



Answer 14 is a and d are true

- Statement a is true, from the regression line. The predicted ATP of youngest son if the ATP of oldest son is 5 is clearly less than 5, based on the regression line.

- As for Statement b, 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.
- Statement c 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 `young.atp` = 0.99 + 0.83 `old.atp`
- Statement d is also 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.

```
lm(young.atp ~ old.atp, data = dat14)
```

```
Call:
lm(formula = young.atp ~ old.atp, data = dat14)

Coefficients:
(Intercept)      old.atp
     0.9867       0.8337
```

```
cor(dat14 %>% select(old.atp, young.atp))
```

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

Results

Item	14
Correct (out of 60)	53
% of Available Points Awarded	> 90

Partial credit was awarded according to the following schedule.

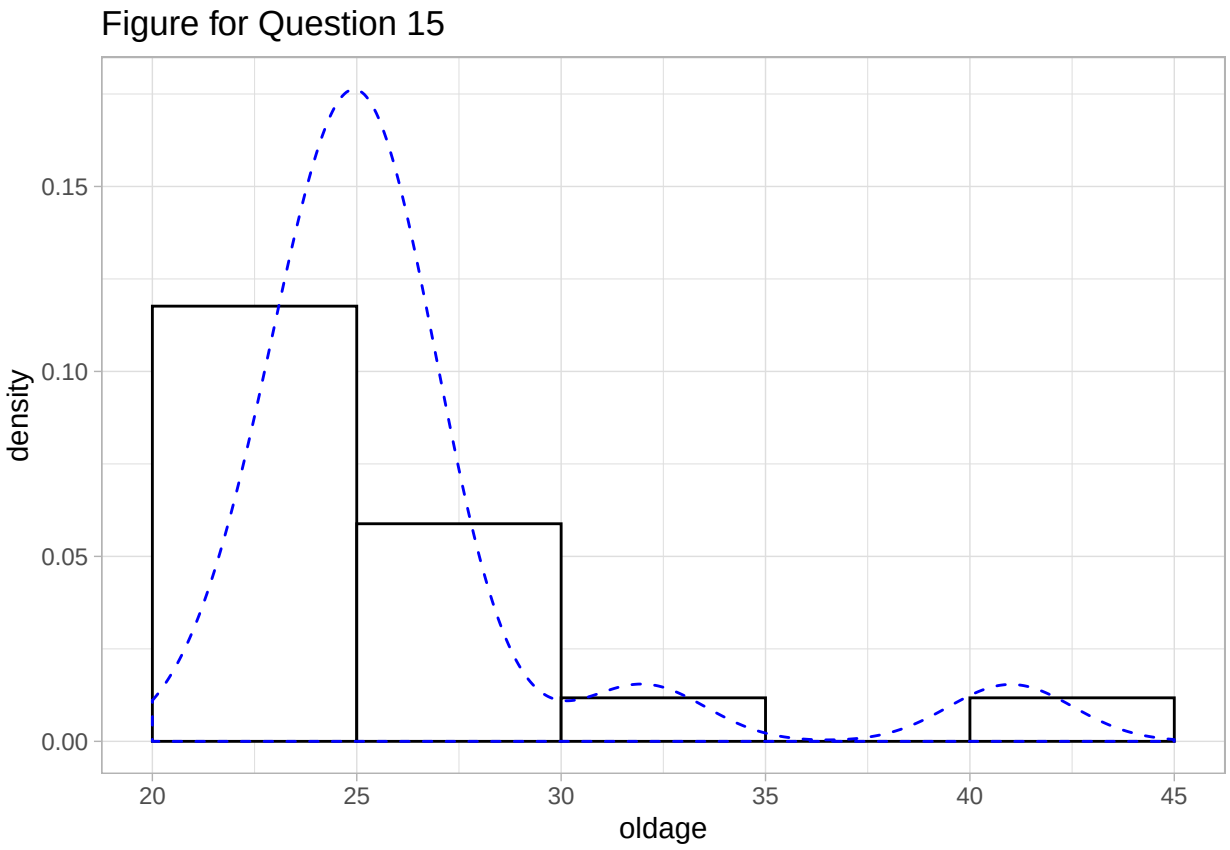
Response	Points Awarded
a, d	4
a	2
a, b, d	2
a, c, d	2
d	2

Question 15

Consider again the study described in Question 14, but now, we'll focus on the ages of the oldest sons. The Figure for Question 15 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. A Normal Q-Q plot of these data would show an S-shape.
b. The ages are symmetric, showing no substantial skew.
c. The mean of the ages is larger than the median age.
d. The range of the data covers somewhere between 15 and 25 years.
e. None of these statements are true

Figure for Question 15



Answer 15 is both c and d.

- These are right-skewed data, according to the histogram.
- Statement a would be true only if the data were symmetric but light-tailed. Not the case here.
- Statement b is also false because the data are not symmetric.
- Statement c is true, because the data are right skewed, which pulls the mean above the median.
- Statement d is also true. 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).

Results

Item	15
Correct (out of 60)	13
% of Available Points Awarded	53

Partial credit was awarded according to the following schedule. The most common response was `c` alone.

Response	Points Awarded
c, d	3
a	0
a, c	1
a, c, d	1.5
c	1.5
d	1.5
e	0

Question 16

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. Province of residence for a group of Canadian citizens.
d. Days between attacks for a patient diagnosed with relapsing-remitting multiple sclerosis.
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.
f. Highest level of education completed (grade school, high school, college, higher than college)

The columns are:

- Quantitative
- Ordinal categorical
- Nominal categorical
- It is impossible to tell

Answers for Question 16 are as listed below:

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

Results

Item	16a	16b	16c	16d	16e	16f
Correct (out of 60)	60	60	60	> 54	60	> 54
% of Available Points Awarded	100	100	100	> 90	100	> 90

You received 0.5 point for each correct response. No partial credit was available.

Setup for Questions 17-19

As part of the materials for the Quiz, you'll find an R Markdown file called `questions17-19_initial.Rmd` and an HTML file called `questions17_19_resultswewant.Rmd`. In order to generate that `resultswewant` file, you need to solve three problems with the `initial` R Markdown file. In Questions 17-19, we ask you to identify (by the line number in the initial file) the three places where a change needs to be made, and specify what that change should be in order to produce the results we want.

Question 17

There are three critical errors in the initial R Markdown file. List the first one by specifying the line number in which it occurs in the initial R Markdown file, then specify how it needs to be fixed.

Answer 17 is a sentence.

In line 20, the tabyl doesn't show both row and column totals. To achieve this, we must change `adorn_totals()` to `adorn_totals(where = c("row", "col"))`. Some people did this by piping together both `adorn_totals("row")` and `adorn_totals("col")`.

Question 18

There are three critical errors in the initial R Markdown file. List the second one by specifying the line number in which it occurs in the initial R Markdown file, then specify how it needs to be fixed.

Answer 18 is a sentence.

In line 25, we repeat the code chunk name look_at_diamonds_data that we used in line 17. We need to change them so that we don't use the same chunk name twice.

Question 19

There are three critical errors in the initial R Markdown file. List the third one by specifying the line number in which it occurs in the initial R Markdown file, then specify how it needs to be fixed.

Answer 19 is a sentence.

In line 30, we failed to include a + after `theme_bw()` which we need in order to get the follow-up `labs` command to do what we want.

Grading Questions 17-19

Results

Item	17-19 tabyl denom	chunk name	add +
Correct (out of 60)	> 54	> 54	> 54

Item	17-19 tabyl denom	chunk name	add +
% of Available Points Awarded	> 90	> 90	> 90

It was not important that you place the three errors in any particular order. You received three points for each error that you correctly identified. If you got the right answers but specified them in the wrong lines, we ignored that.

Bonus Question 19x. (optional: 2 points of extra credit)

There is actually a fourth problem in the initial R Markdown file, which doesn't need to be corrected to get the `resultswewant` HTML file, but which should be corrected anyway. Specify its line number in the initial file, and then specify how it needs to be fixed.

Answer to Bonus Question 19x

In line 10, we have `library(ggplot2)` and in line 12, we have `library(tidyverse)` but `ggplot2` is part of the core tidyverse. We should remove line 10.

Grading Question 19x

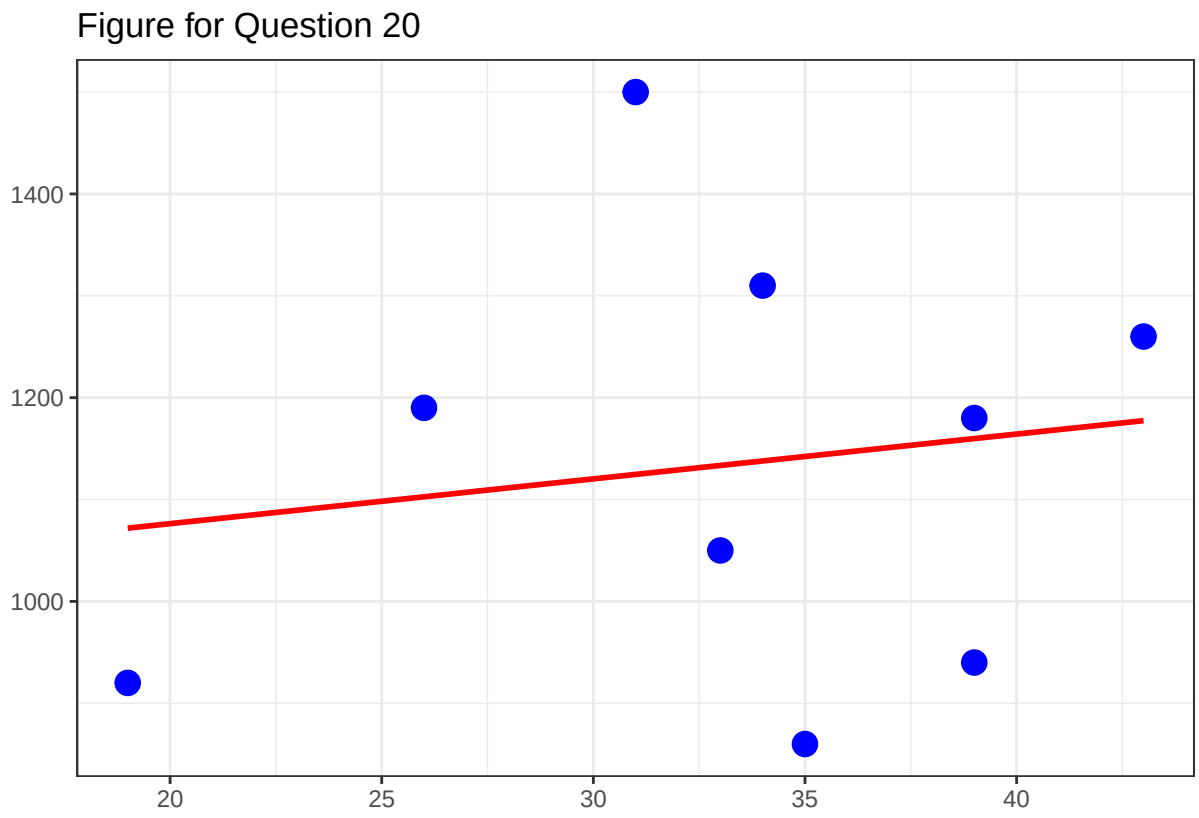
Item	19X (BONUS)
Correct (out of 60)	5
% of Available Points Awarded	10

- If you tried to remove the tidyverse instead of removing `ggplot2` that would have created problems, so no points there.
- The only way to get 1 point of partial credit here is if you named something that you didn't name in the actual Questions 17-19 that should have been there.

Question 20 (4 points)

Fast food is often high in both fat and sodium. But are the two related? The scatter plot shown in the Figure for Question 20 describes the fat (in g) and sodium (in mg) contents of nine 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 20



Answer 20 is Add axis titles.

The correct response is to label the axes.

Results

Item	21
Correct (out of 60)	51
% of Available Points Awarded	86

Grading Notes

- If your answer included labeling the axes, you got 4 points, unless you also wrote something that is incorrect.
  - We gave full credit whether or not you specified the actual axis titles you wanted, or if you specified that units would be helpful, or if you also requested other things that were at least reasonable.
  - Merely saying the labeling is unintutitive got you 2 points.
- If your answer did not include labeling the axes, you did not receive credit.

Question 21

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 <- tibble(rnorm(n = 25, mean = 25, sd = 4))
c. scores <- rep(rnorm(mean = 25, sd = 4), 1000)
d. scores %>% rnorm(n = 1000, mean = 25, sd = 4)
e. scores <- tibble(y <- rnorm(1000, mean = 25, sd = 4))
f. None of these commands.

Answer to Question 21 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)
```

or perhaps you could put this within a tibble called `dat21` as:

```
dat21 <- tibble(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` produces 25 observations rather than 1000, and puts them in a tibble called scores, rather than a variable called scores.
- `c` will produce an error since there's no n value in your rnorm call, among other things.
- `d` uses the pipe %>% rather than the assignment operator <-
- `e` will put the data into a variable called "y <- rnorm(1000, mean = 25, sd = 4)" within a tibble called scores, and while I suppose that's a little closer, it's still not right.

Results

Item	21
Correct (out of 60)	10
% of Available Points Awarded	39

Unsurprisingly, people are loathe to choose "none of the above."

- I was probably overly generous in giving 1 point of partial credit for `e` as a response, since that's wrong, too. `e` alone was the most common response.
- If you gave `e` as well as other responses, no credit.

Question 22 (4 points)

I have provided you with a data set called `dat22.csv`. After you import that into R as a tibble called `item22`, the result should contain a variable called `apgar5` that contains scores on the APGAR scale at five minutes for 130 infants, although 4 of the values are listed as NA.

You wish to obtain the standard deviation of the APGAR scores in the `item22` tibble. 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 126 infants not marked as NA. (CHECK ALL THAT APPLY.)

a. mosaic::favstats(~ apgar5, data = item22)
b. summary(item22)
c. item22 %$% sd(apgar5)
d. item22 %>% summarize(sd(apgar5, na.rm = TRUE))
e. item22 %>% filter(complete.cases(apgar5)) %>% summarize(sd = sd(apgar5))
f. item22 %>% select(complete.cases(apgar5)) %>% summarize(sd = sd(apgar5))
g. None of these will produce the correct value.

Answer to Question 22 is a, d and e

The `item22` tibble looks like what we want.

```
item22 <- read_csv("data/dat22.csv")

item22
```

```
# A tibble: 130 x 2
   subject apgar5
     <dbl>  <dbl>
 1       1      8
 2       2      7
 3       3     NA
 4       4      7
 5       5      9
 6       6      7
 7       7      8
 8       8      5
 9       9      7
10      10      7
# ... with 120 more rows
```

```
item22 %>% tabyl(apgar5)
```

```
 apgar5  n     percent valid_percent
      4  1 0.007692308   0.007936508
      5  3 0.023076923   0.023809524
      6 14 0.107692308   0.111111111
      7 24 0.184615385   0.190476190
      8 35 0.269230769   0.277777778
      9 37 0.284615385   0.293650794
     10 12 0.092307692   0.095238095
     NA  4 0.030769231            NA
```

Now, what works to provide the correct standard deviation (1.3) for the 126 "apgar5" values?

- Statement a works.

```
mosaic::favstats(~ apgar5, data = item22)
```

```
 min Q1 median Q3 max     mean       sd   n missing
   4  7      8  9  10 7.968254 1.289567 126       4
```

- Statement b doesn't work because the `summary` function doesn't present the standard deviation.

```
summary(item22)
```

```
    subject           apgar5
 Min.   :  1.00   Min.   : 4.000
 1st Qu.: 33.25   1st Qu.: 7.000
 Median : 65.50   Median : 8.000
 Mean   : 65.50   Mean   : 7.968
 3rd Qu.: 97.75   3rd Qu.: 9.000
 Max.   :130.00   Max.   :10.000
                  NA's   :4
```

- Statement c doesn't work because `apgar5` has missing (NA) values within the `item22` tibble, so the result this gives is NA. You'd have to include na.rm=TRUE to make it work, for example, or filter to only the complete cases within the `item22` tibble.

```
item22 %$% sd(apgar5)
```

```
[1] NA
```

- Statement d works.

```
item22 %>% summarize(sd(apgar5, na.rm = TRUE))
```

```
# A tibble: 1 x 1
  `sd(apgar5, na.rm = TRUE)`
                       <dbl>
1                       1.29
```

- Statement e works.

```
item22 %>% filter(complete.cases(apgar5)) %>% summarize(sd = sd(apgar5))
```

```
# A tibble: 1 x 1
     sd
  <dbl>
1  1.29
```

- Statement f 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`.

```
item22 %>% select(complete.cases(apgar5)) %>% summarize(sd = sd(apgar5))
```

Note that this last one throws an error message, so I didn't execute it here.

Results

Item	22
Correct (out of 60)	46
% of Available Points Awarded	88

Partial credit was awarded according to the following schedule.

Response	Points Awarded
a, d, e	4
a, c, d, e	2
a, d	2
a, e	2
d, e	2
e	1

Question 23

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 50 children, gathering each child's weight percentile and the number of cigarettes smoked per day in the child's home. For the 50 children in the new study, the Pearson correlation coefficient between weight percentile and cigarettes smoked was reported as r = -0.55. 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 50 children. Which one 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 23 is c

If r = -0.55, then r^2 will be 30.25% (so the model accounts for 30.25% of variation), and the slope will be negative, because the least squares regression line's slope and the Pearson correlation coefficient are defined so that they must always have the same sign.

Results

Item	23
Correct (out of 60)	45
% of Available Points Awarded	75

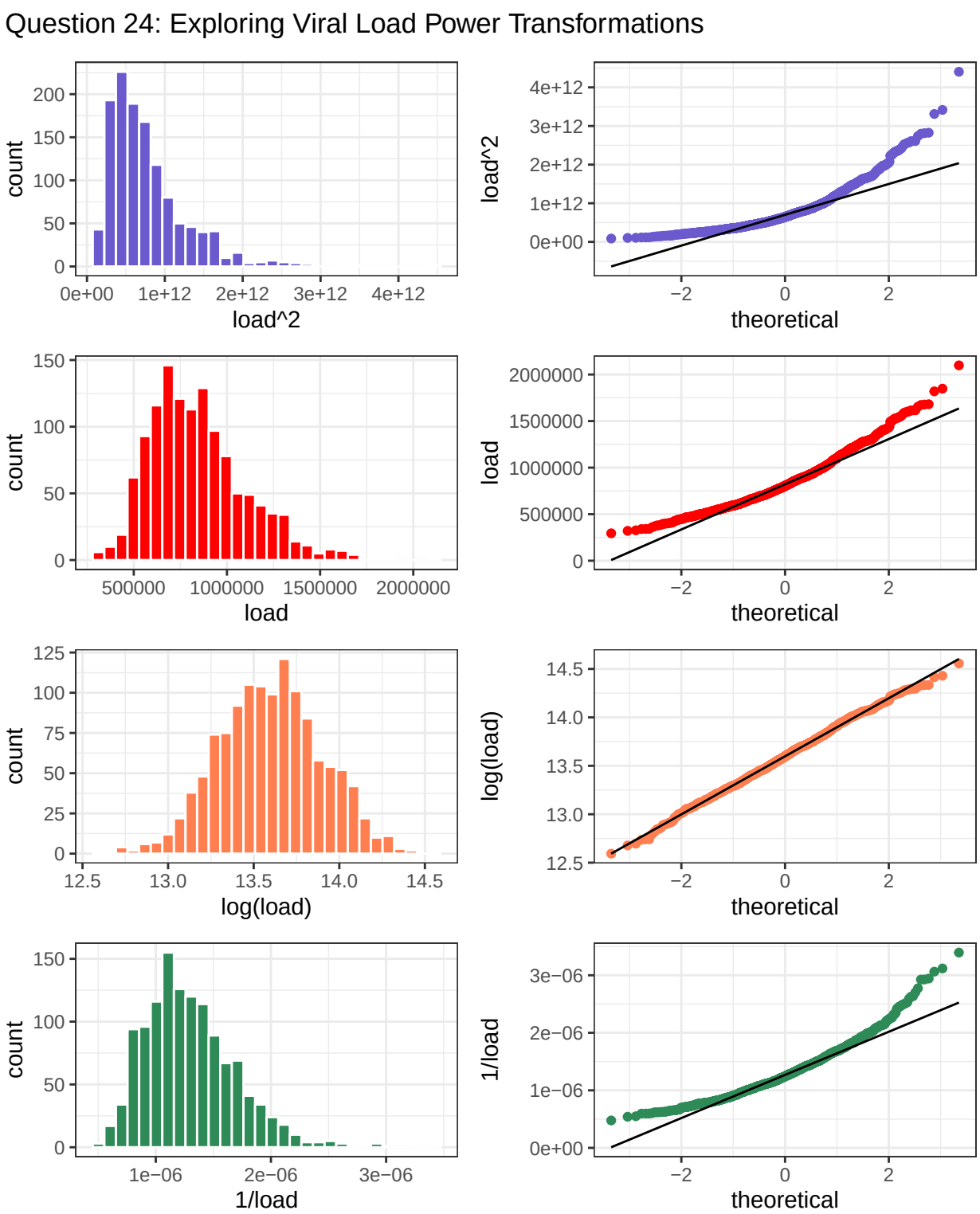
There was no partial credit available on this question.

Question 24

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 Figure for Question 24. 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 24



Answer 24 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.

Results

Item	24
Correct (out of 60)	> 54
% of Available Points Awarded	> 90

There was no partial credit available on this question.

Question 25

A new sample of 350 subjects ages 35-59 from the NHANES data generates the Table for Question 25, 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 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 25

	HealthGen					
Marijuana	Excellent	Vgood	Good	Fair	Poor	Total
No	12	44	58	21	2	137
Yes	20	78	77	36	2	213
Total	32	122	135	57	4	350

Answer 25 is d

We could just do the math.

- In the "Yes" group, we have 20 + 78 + 77 = 175 people in the three best health groups, and that's out of a total of 213 people in the "Yes" group, so that's 82.2%.
- In the "No" group, we have 12 + 44 + 58 = 114 people in the three best health groups, and that's out of a total of 137 people in the "No" group, so that's 83.2%.

Or, we could get some help. Here are the health category percentages, within each marijuana group.

```
dat25 %>% tabyl(Marijuana, HealthGen) %>%
  adorn_percentages(denominator = "row") %>%
  adorn_pct_formatting()
```

```
 Marijuana Excellent Vgood  Good  Fair Poor
        No      8.8% 32.1% 42.3% 15.3% 1.5%
       Yes      9.4% 36.6% 36.2% 16.9% 0.9%
```

- So, in the "No" group, we have 8.8 + 32.1 + 42.3 = 83.2 percent in the three healthiest categories.
- In the "Yes" group, we have 9.4 + 36.6 + 36.2 = 82.2 percent in the three healthiest categories.

In either case, that is a difference of 1.0 percentage point, favoring the "No" group. That's choice **d**.

Results

Item	25
Correct (out of 60)	42
% of Available Points Awarded	70

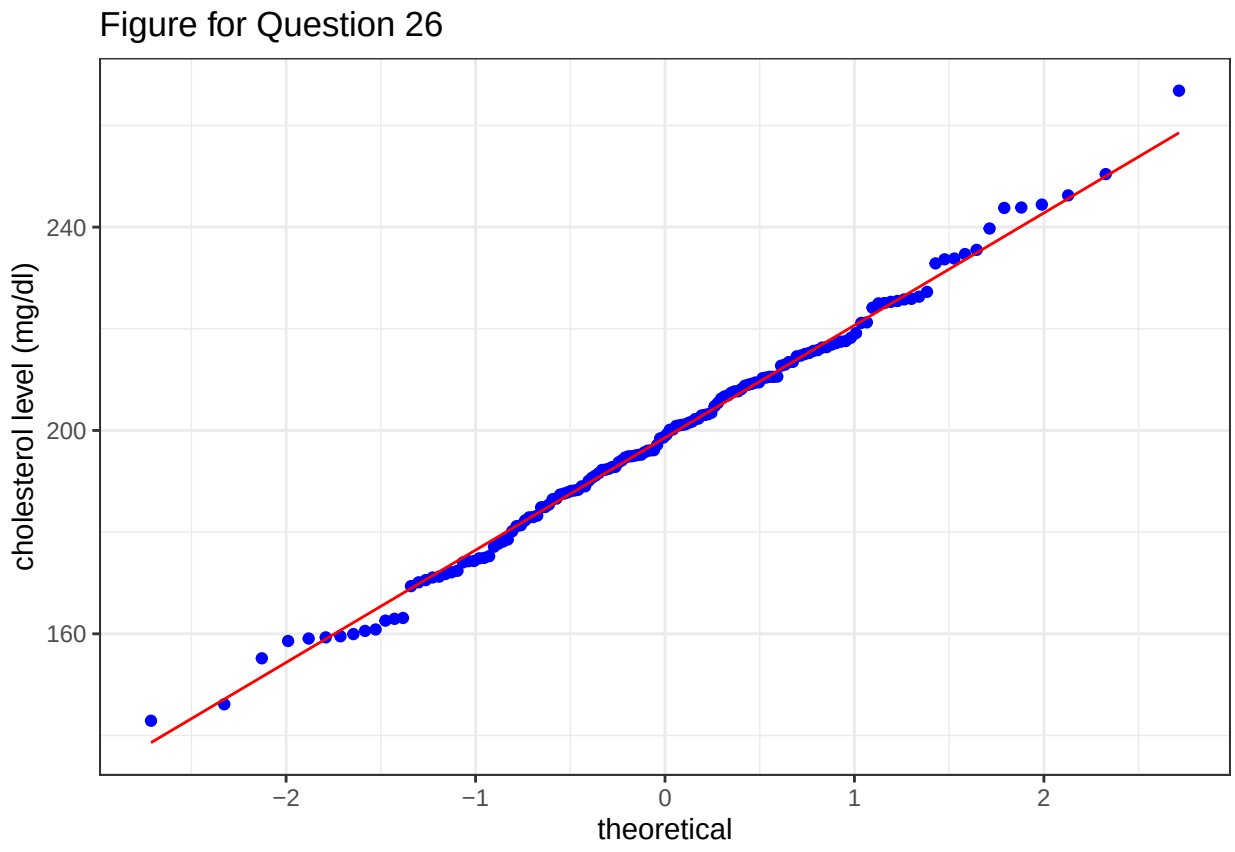
There was no partial credit available on this question.

Question 26

The Figure for Question 26 is a Normal Q-Q plot of cholesterol levels (in mg/dl) for 150 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 250 mg/dl and a standard deviation of approximately 10 mg/dl.
d. Approximately Normally distributed, with a mean of approximately 200 mg/dl and a standard deviation of approximately 25 mg/dl.
e. Approximately Normally distributed, with a mean of approximately 200 mg/dl and a standard deviation of approximately 10 mg/dl.
f. Not approximately Normally distributed, but instead substantially skewed to the left.
g. Not approximately Normally distributed, but instead substantially skewed to the right.

Figure for Question 26



Answer 26 is d

- 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 re-calibrate your expectations for a Normal Q-Q plot.

Results

Item	26
Correct (out of 60)	50
% of Available Points Awarded	83

There was no partial credit available on this question.

Question 27

Choose the five number summary (minimum, Q1, median, Q3 and maximum) that matches the stem-and-leaf plot of LDL cholesterol levels shown in the Figure for Question 27.

a. Min: 53 Q1: 100 Median: 135 Q3: 155 Max: 241
b. Min: 53 Q1: 100 Median: 131 Q3: 158 Max: 241
c. Min: 53 Q1: 100 Median: 131 Q3: 155 Max: 241
d. Min: 53 Q1: 100 Median: 122 Q3: 155 Max: 241
e. Min: 53 Q1: 100 Median: 135 Q3: 158 Max: 241
f. Min: 53 Q1: 100 Median: 122 Q3: 158 Max: 241

Figure for Question 27

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

 5 | 33
 6 |
 7 | 1
 8 | 8
 9 | 39
10 | 078
11 | 07
12 | 2
13 | 15
14 | 1456
15 | 58
16 |
17 | 08
18 |
19 | 9
20 |
21 | 1
22 |
23 |
24 | 1
```

Answer 27 is c

```
fivenum(dat27$ldl)
```

```
[1]  53 100 131 155 241
```

or

```
summary(dat27)
```

```
      rand            ldl
 Min.   : 1   Min.   : 53
 1st Qu.: 7   1st Qu.:100
 Median :13   Median :131
 Mean   :13   Mean   :131
```

```
 3rd Qu.:19   3rd Qu.:155
 Max.   :25   Max.   :241
```

Results

Item	27
Correct (out of 60)	> 54
% of Available Points Awarded	> 90

There was no partial credit available on this question.

Question 28 (4 points)

Suppose you have collected data into a tibble in R called `dat28`. The `dat28` data come from a cohort study to look at the impact of exposure to an industrial solvent (stored in the `solvent` variable: a factor taking on the values "none", "moderate" or "profound") on the probability of a bladder cancer diagnosis (stored as either yes or no in the `diagnosis` variable.)

Provide a single line of R code to obtain an appropriate summary of the association between these variables. You may include at most one pipe in your response.

Answer 28 is `dat28 %>% tabyl(solvent, diagnosis)`

The answer I was looking for is `dat28 %>% tabyl(solvent, diagnosis)`, which produces a cross-tabulation with the `solvent` information in the rows and the `diagnosis` information in the columns.

Here, I'll make up a data set to show you.

```
dat28 <- tibble(
  id = c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10),
  solvent = c("none", "none", "none", "moderate", "moderate", "moderate", "profound", "profound", "profo
  diagnosis = c("yes", "no", "no", "no", "no", "yes", "no", "yes", "yes", "yes")
)

dat28 %>% tabyl(solvent, diagnosis)
```

```
  solvent no yes
 moderate  2   1
     none  2   1
 profound  1   3
```

Results

Item	29
Correct (out of 60)	32
% of Available Points Awarded	55

Grading Notes

- Some people put the diagnosis in the rows, some in the columns. Either was OK.
- Some people used tools we haven't studied at all (or at least not in detail), like `xtabs` or `ftable`, to do this work. If you did that *correctly* so that the result produces a cross-classification table, you got full credit. So, for instance, `addmargins(xtabs(~ solvent + diagnosis, data = dat28))` got full credit.
- A call to `count` can work here, too, although it doesn't produce a cross-tabulation. I gave full credit for this, although it's clearly not as useful as `tabyl` would be.
- As long as you didn't make a mistake, it didn't matter whether you asked for row totals, column totals, both or neither in your tibble.
- Some people produced percentages or proportions instead of counts. Less useful in some settings, but still OK for this purpose.
- Some people tried to use the pipe in functions that don't take it without "exploding" the variables, so for instance: `dat28 %>% table(solvent, diagnosis)` doesn't work while `dat28 %$% table(solvent, diagnosis)` would have worked. I gave 1 point of partial credit in that instance.

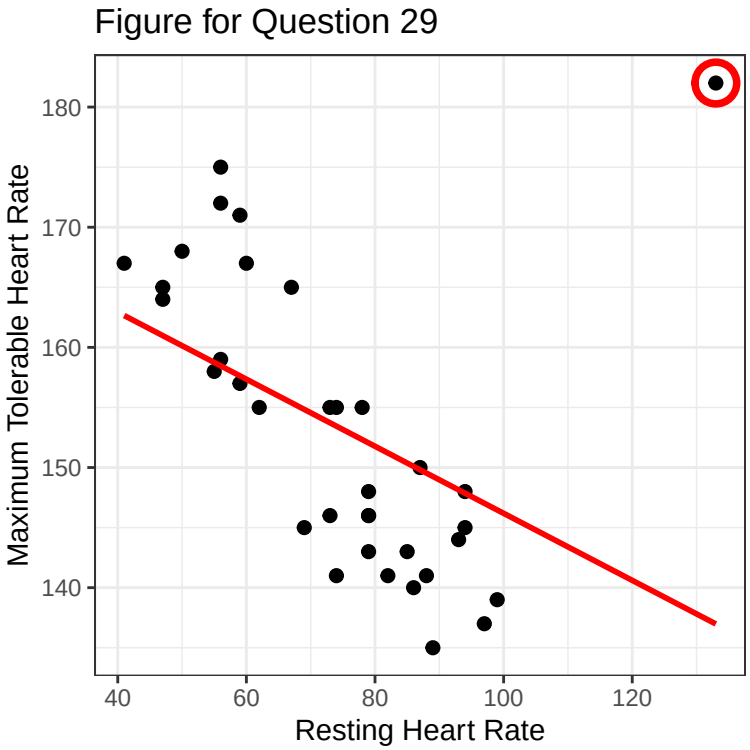
- Some people didn't specify the tibble, writing, for instance, `tabyl(solvent, diagnosis)` alone, and thus got 1 point of partial credit.
- You needed to use the actual variable names: solvent and diagnosis. If you didn't, you got at most 1 point.
- Some people felt like they had to do a chi-squared test. Without presenting the cross-tabulation, that didn't get you anywhere useful, or any points. If you included the cross-tabulation, then I ignored the chi-square test part.
- A call to the `cor` function or to `favstats` is incorrect, as neither solvent nor diagnosis is a quantitative variable. Turning one or more of the variables into a number first isn't going to help, either.
- Some people tried to use `group_by` and `summarize` but that's not much help with categorical diagnosis data.
- Some people produced graphs, but a scatterplot doesn't do you any good here.
- Some people tried to fit linear regression models. Not the idea, and it wouldn't work anyway.
- Some people only summarized one of the variables.
- One person left this blank, which was a shame.

Question 29

The scatterplot shown in the Figure for Question 29 displays data on resting heart rate and maximum tolerable heart rate for 35 subjects in a research study. Subject 11, whose data are circled in red, has a resting heart rate of 133 and a maximum tolerable rate of 182. If the scatterplot was redrawn including only the other 34 subjects, the Pearson correlation coefficient would do what?

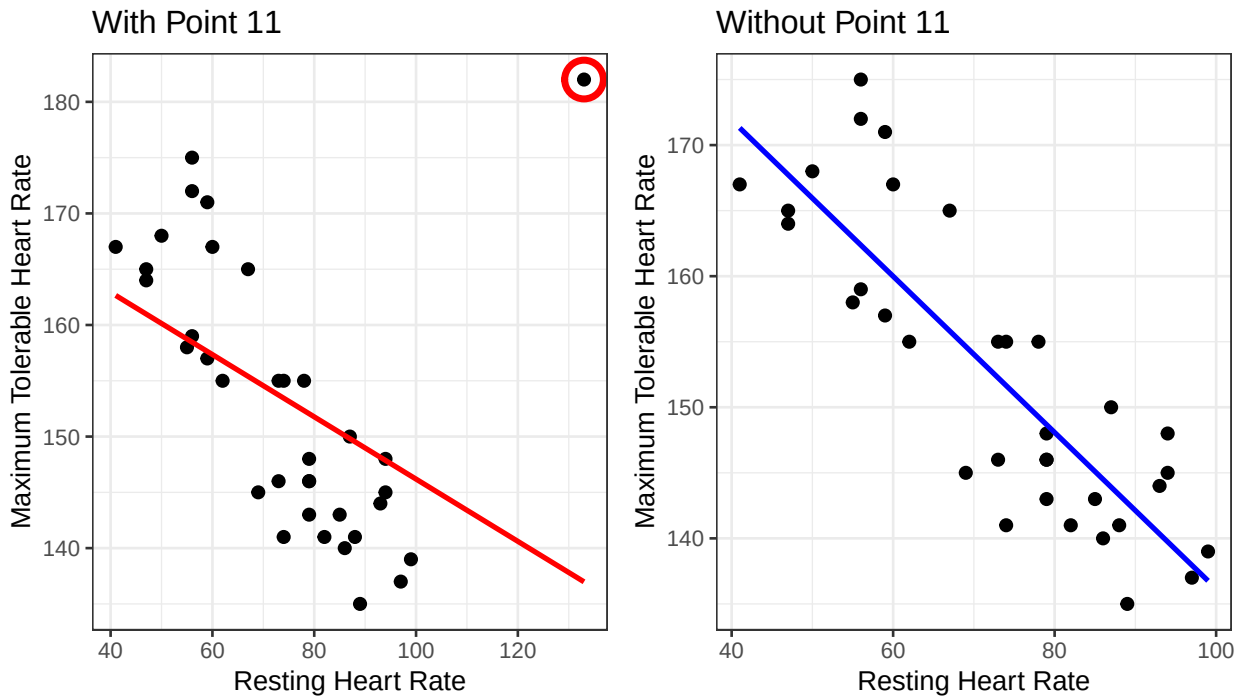
a. decrease
b. increase
c. remain unchanged
d. It is impossible to tell from the information provided.

Figure for Question 29



Answer 29 is a

Removal of the outlier would cause the association to be stronger, that is to say, it would make the Pearson correlation coefficient (which is already negative) move closer to -1, and thus decrease.



You didn't have the data, so you couldn't figure this out on your own using R (you had instead to look at the plot and think through what would happen), but in fact the Pearson correlation of the original data set with 35 observations is -0.43, but after we remove point 11 from the data, the Pearson correlation of the remaining 34 observations turns out to be -0.845.

```
q29 %$% cor(resting, max.tolerance)
```

```
[1] -0.4300022
```

```
q29 %>% slice(-11) %$%
  cor(resting, max.tolerance)
```

```
[1] -0.844986
```

Results

Item	29
Correct (out of 60)	40
% of Available Points Awarded	67

There was no partial credit available on this question. I am aware that people twisted themselves into all sorts of knots in responding to this question. I encourage you to try not to do that.

Question 30 (4 points)

According to the *Elements of Data Analytic Style*, which of the following elements belong in a proper written data analysis? (CHECK ALL THAT APPLY.)

a. An introduction or motivation.
b. A description of the statistical models used.
c. Conclusions including potential problems.
d. A link to the code used to produce the analysis, including all figures and tables.
e. References
f. A meaningful title that clearly conveys the key research question.
g. Specification of the main results on the scientific scale of interest.
h. Measures of uncertainty (like confidence intervals) alongside point estimates.
i. Reports of potential problems with the analysis.

Answer 30 is all 9 of them.

See Chapter 9 of Leek's book.

Results

Item	30
Correct (out of 60)	18
% of Available Points Awarded	70

Partial credit was awarded according to the following schedule.

Response	Points Awarded
All 9	4
8 out of 9	3
7 out of 9	2
6 out of 9	1
5 or fewer1	0

The most common option people left out was `d`.

Overall Results

Add up your points on the 30 items, then add 5 points to that raw total. That's your score.

Scores on the Test	Interpretation
91 - 99	A
85 - 90.5	A-
81 - 84.5	B+
75 - 80.5	B
70 - 74.5	B-
60 - 69.5	C

Scores on the Test	Interpretation
below 60	see Dr. Love