431 Fall 2020 Quiz 1

Thomas E. Love

Version 2020-10-02 17:06:48. Deadline is Noon 2020-10-07

Instructions

There are 29 questions on this Quiz. It is to your advantage to answer all 29 Questions. Your score is based on the number of correct responses, so there's no chance a blank response will be correct, and a guess might be, so you should definitely answer all of the questions.

0.1 The Google Form Answer Sheet

All of your answers should be placed in the Google Form Answer Sheet, located at http://bit.ly/431-2020-quiz1-answer-sheet. All of your answers must be submitted through the Google Form by noon on Wednesday 2020-10-07, without exception. The form will close at that time, and no extensions will be made available, so do not wait until Wednesday to submit. We will not accept any responses except through the Google Form.

The Google Form will contain places to provide your responses to each question, and a final affirmation where you'll type in your name to tell us that you followed the rules for the Quiz. You must complete that affirmation and then submit your results. When you submit your results (in the same way you submit a Minute Paper) you will receive an email copy of your submission, with a link that will allow you to edit your results.

If you wish to work on some of the quiz and then return later, you can do this by [1] completing the final question (the affirmation) which asks you to type in your full name, and then [2] submitting the quiz. You will then receive a link at your CWRU email which will allow you to return to the quiz as often as you like without losing your progress.

0.2 The Data Sets

I have provided three data sets (called `algae.csv`, `oscar.csv` and `tobacco_boys.csv`) that are mentioned in the Quiz. They may be helpful to you.

0.3 Getting Help

This is an open book, open notes quiz. You are welcome to consult the materials provided on the course website and that we've been reading in the class, but you are not allowed to discuss the questions on this quiz with anyone other than Professor Love and the teaching assistants. You will be required to complete a short affirmation that you have obeyed these rules as part of submitting the Q

If you need clarification on a Quiz question, you have exactly two ways of getting help:

1. You can ask your question in a private post on Piazza to the instructors. (This is the only kind of post you will be able to make on Piazza during the Quiz.)
2. You can ask your question via email to **431-help at case dot edu**.

During the Quiz period (5 PM 2020-10-02 through noon 2020-10-07) we will not answer questions about the Quiz except through the two approaches listed above. We promise to respond to all questions received before 9 AM on 2020-10-07 in a timely fashion.

A few cautions:

- Specific questions are more likely to get helpful answers.
- We will not review your code or your English for you.
- We will not tell you if your answer is correct, or if it is complete.
- We will email all students if we find an error in the Quiz that needs fixing.

0.3.1 When Should I ask for help?

We recommend the following process.

- If you encounter a tough question, skip it, and build up your confidence by tackling other questions.
- When you return to the tough question, spend no more than 10-15 minutes on it. If you still don't have it, take a break (not just to do other questions) but an actual break.
- When you return to the question, it may be much clearer to you. If so, great. If not, spend 5-10 minutes on it, at most, and if you are still stuck, ask us for help.
- This is not to say that you cannot ask us sooner than this, but you should **never, ever** spend more than 20 minutes on any question without asking for help.

0.4 Scoring and Timing

All questions are worth either 3 or 4 points, adding to a total of 100 points. The 13 questions which are worth 4 points each (specifically, Questions 1, 4, 6, 10, 12, 13, 14, 15, 20, 24, 25, 26 and 29) are marked as such in the Quiz. The questions are not in any particular order, and range in difficulty from "things I expect everyone to get right" to "things that are deliberately tricky".

The Quiz is meant to take 4 hours. I expect most students will take 3-5 hours, and some will take as little as 2 or as many as 8. It is not a good idea to spend a long time on any one question.

0.5 Other Issues

Occasionally, we ask you to provide a single line of code. If not otherwise specified, a single line of code in response can contain at most two pipes, 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.

R Packages Dr. Love Used To Build This

This doesn't mean you need to use all of these packages.

```
library(broom)
library(car)
library(ggrepel)
library(janitor)
library(knitr)
library(magrittr)
library(patchwork)
library(tidyverse)
```

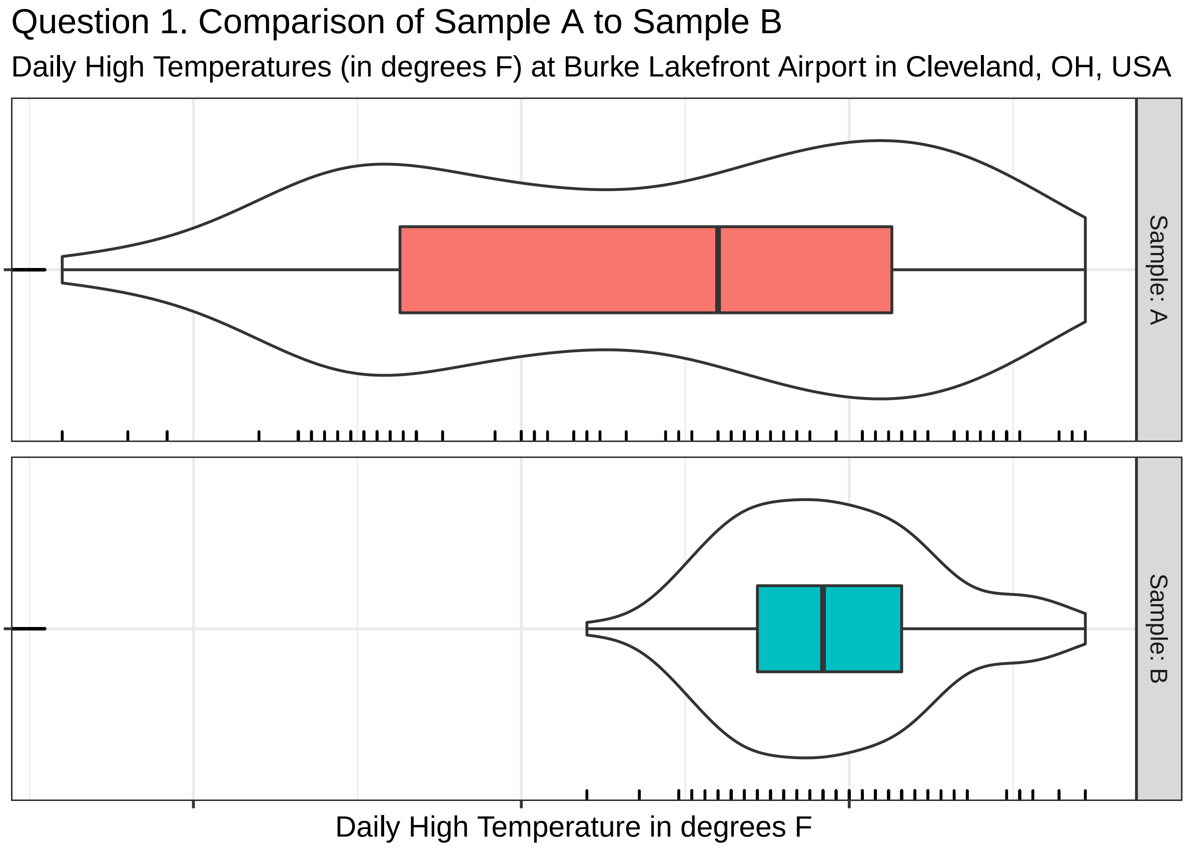
```r
theme_set(theme_bw())
```

1 Question 1. (4 points)

The plot for Question 1 shows the high daily temperatures (in degrees Fahrenheit) measured at Burke Lakefront Airport in Cleveland, Ohio in two groups of dates, drawn from the past few years.

- One of the samples was formed from a random selection of 100 dates in the month of September.
- The other sample includes a random selection of 100 dates from the entire year.

Unfortunately, the x-axis (which was the same for each subplot) was left unlabeled, **but the missing x-axis labels are the same** for each of the two samples of data. The plot below provides some evidence regarding the distributions of the two samples.



Which of the following statements are true?

```
I. Sample A describes the data gathered only in September.
II. The interquartile range in Sample A is wider than that of Sample B.
III. Sample A would be less accurately modeled using a Normal distribution than Sample B.
```

a. I only
b. II only
c. III only
d. I and II
e. I and III
f. II and III
g. All three statements
h. None of the three statements

2 Question 2.

A regression model performed to predict selling prices of houses found the equation

`Price = 196283 + 54.3 Area + 0.724 Lotsize - 6592 Age`

where `Price` is in dollars, `Area` is in square feet, `Lotsize` is in square feet and `Age` is in years. The data included 250 houses, and the R-squared value is 84%.

Which of the interpretations listed below is most correct?

a. This model fits 84% of the data points exactly.
b. Each year a house ages it is worth $6592 less than it was the year before.
c. Every dollar in price means `Lotsize` increases by 0.724 square feet.
d. The correlation between predicted `Price` using this model and actual `Price` is 0.84.
e. Every extra square foot of `Area` is associated with an additional $54.30 in average price, for houses with a given `Lotsize` and `Age`.

3 Question 3.

The process of inductive inference, as described in *The Art of Statistics*, requires us to think hard about how we move from looking at the raw data to making general claims about the target population. Consider the following principles of effective measurement in this context.

```
I. We want to actually measure what we really want to measure
   without introducing systematic bias.

II. We want to sample at random whenever possible from the
    available subjects we are trying to make inferences about.

III. We want to use measures that give us a good chance of getting
     a similar result in a new study using the same measures.
```

Each of the principles listed above is associated primarily with a particular step in the process of building inductive inference. Identify the step in the process associated with each of the statements above.

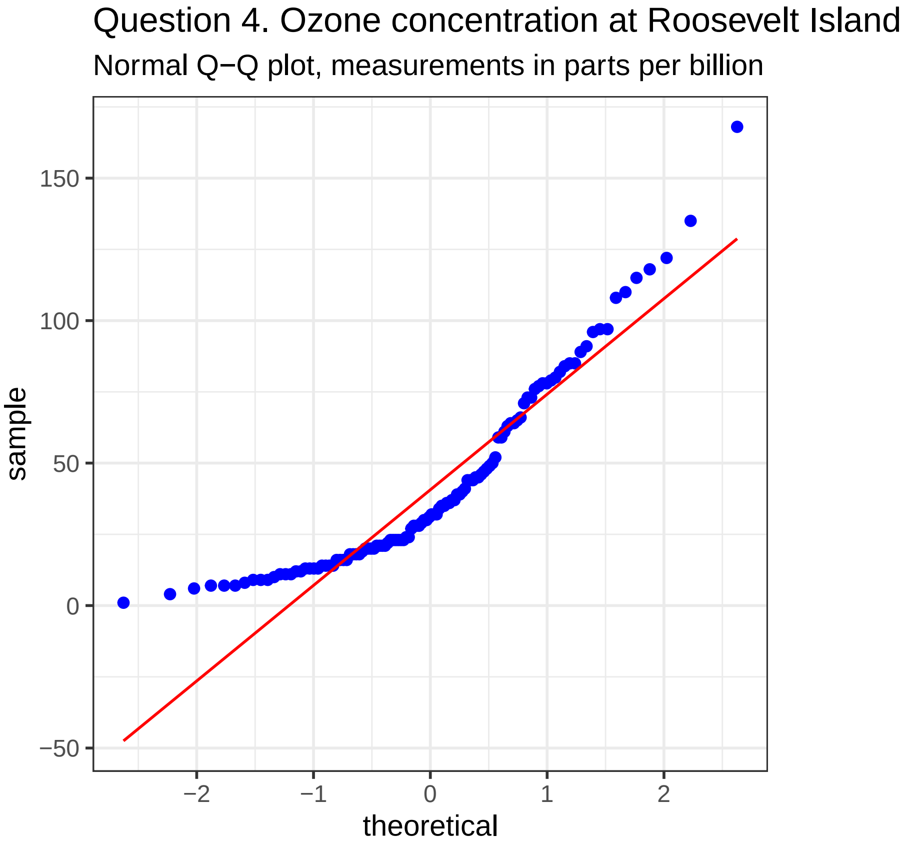
a. Moving from the raw data to the sample
b. Moving from the sample to the study population
c. Moving from the study population to the target population

4 Question 4. (4 points)

Ozone is an important pollutant that causes respiratory discomfort, triggers asthma attacks, and may increase the risk of developing asthma. It is produced from various components of car exhaust by chemical reactions in the air, powered by sunlight. These chemical reactions proceed faster at higher temperatures.

Predictions of ozone concentrations from forecast weather conditions are useful for public health purposes (ozone alerts). Statistical models of ozone levels may also be useful for validating physical/chemical models.

The plot describes measurements from the summer of 1973 in New York City, specifically, the mean concentration of Ozone (in parts per billion) at Roosevelt Island for the period of 1 to 3 PM each day.



Which of these descriptions best fits the Ozone concentrations?

a. Approximately Normally distributed
b. Essentially symmetric, but with outliers
c. Substantially right skewed
d. Substantially left skewed
e. It is impossible to tell from the information provided.

5 Question 5.

Based on the plot in Question 4, consider the following statements about the distribution of Ozone concentrations.

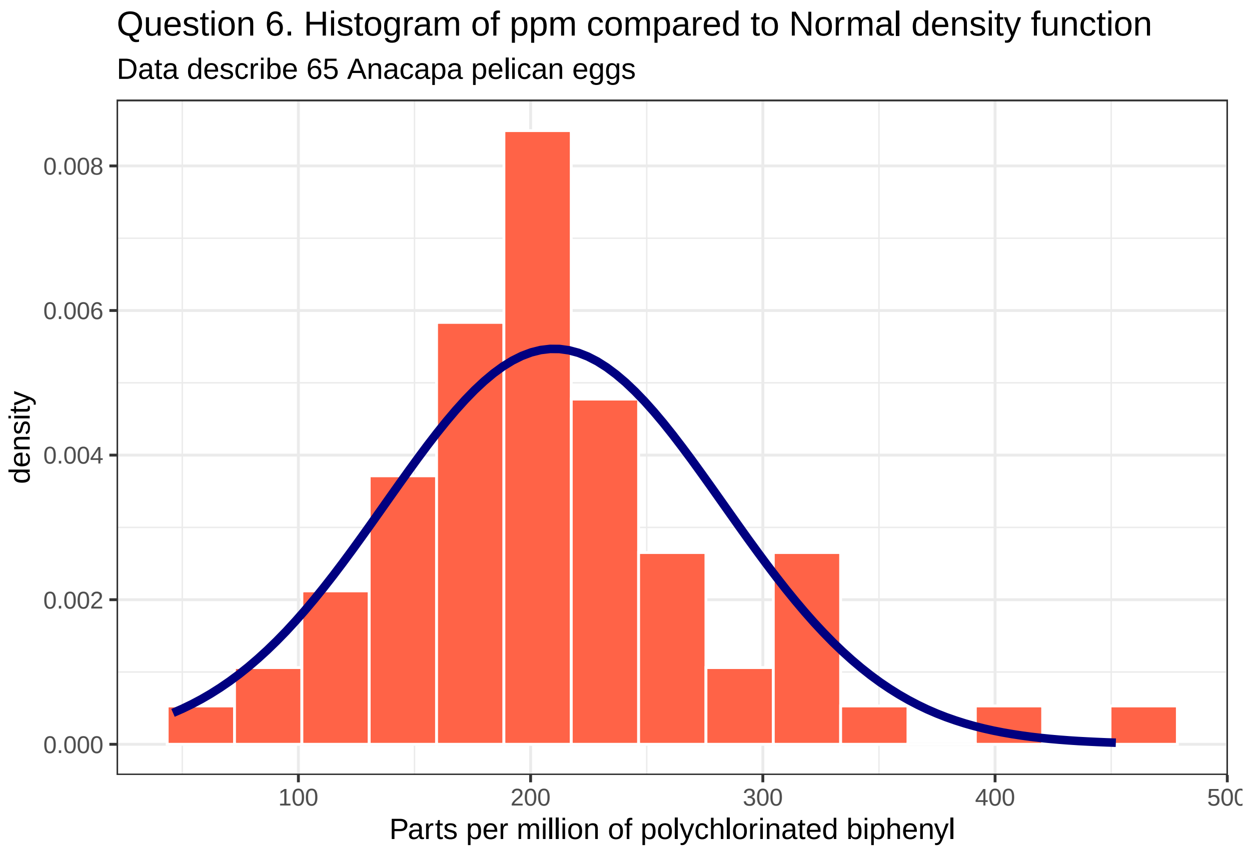
```
I. The mean is below 50 parts per billion.
II. The mean is above 50 parts per billion.
III. The IQR is below 75 parts per billion.
IV. The IQR is above 75 parts per billion.
```

Which of these statements are true?

a. I and III
b. I and IV
c. II and III
d. II and IV
e. It is impossible to tell from the information provided.

6 Question 6. (4 points)

The data for this Question represent the concentration in parts per million of PCB (polychlorinated biphenyl, an industrial pollutant) for 65 Anacapa pelican eggs. The tibble containing the data is called `pelican` and the variable of interest is called `ppm`.



Here are eight lines of code. Note that Dr. Love definitely used lines 1, 2 and 8 in his code. He also used some of the other lines (lines 3-7) but not all of them.

```
1 pelican <- read_csv("data/pelican.csv")

2 ggplot(pelican, aes(x = ppm)) +
3   geom_density(col = "navy", lwd = 1.5) +
4   geom_histogram(aes(y = stat(density)), bins = 15, fill = "tomato", col = "white") +
5   geom_histogram(bins = 15, fill = "tomato", col = "white") +
6   stat_function(fun = dnorm,
                  args = list(mean = mean(pelican$ppm), sd = sd(pelican$ppm)),
                  col = "navy", lwd = 1.5) +
7   coord_flip() +
8   labs(title = "Question 6. Histogram of ppm compared to Normal density function",
         subtitle = "Data describe 65 Anacapa pelican eggs",
         x = "Parts per million of polychlorinated biphenyl")
```

Question 6 continues on the next page.

Continuation of Question 6

Please select each of the line numbers that should be REMOVED from the code in order to create the Question 6 plot. (YOU MAY SELECT MORE THAN ONE OPTION.)

a. Line 3
b. Line 4
c. Line 5
d. Line 6
e. Line 7

7 Question 7.

Suppose you are interested in how effectively shell thickness might be used to predict the concentration of environmental pollutants, in a setting like the study developed in Question 6. Which variable should go on the vertical (Y) axis of your scatterplot to display and model this association?

a. the concentration in parts per million of PCB
b. the thickness in micrometers of the egg's shell
c. the egg identification number (1-65)
d. It doesn't matter.
e. It is impossible to tell from the information provided.

8 Question 8.

In this question, we consider data describing the age at onset (in years) for 17 women with a diagnosis of multiple sclerosis. The oldest age at onset was 44 years. The stem-and-leaf display shows the data for the first 17 subjects.

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

  1 | 46788889
  2 | 0367
  3 | 239
  4 | 24
```

If the next subject added to the data is 28 years of age, which of the following values will decrease, as a result?

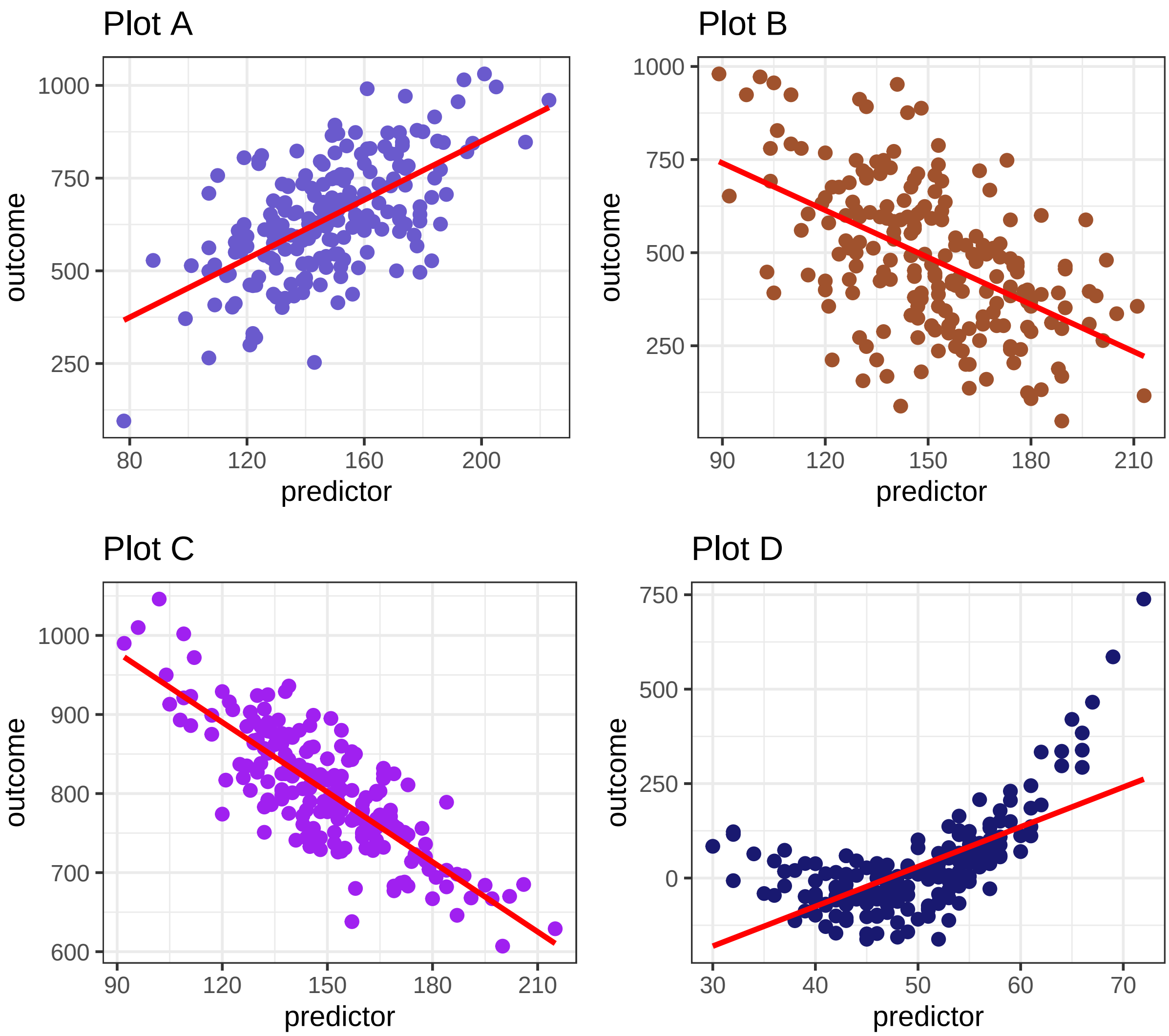
I. The mean
II. The standard deviation
III. The median

a. I only
b. II only
c. III only
d. I and II
e. I and III
f. II and III
g. All three statements
h. None of the three statements

9 Question 9.

Consider the four scatterplots provided for Question 9.

Plots for Question 9



Plot A

Plot B

Plot C

Plot D

Which of the four scatterplots provided for Question 9 is associated with a linear model for `outcome` using `predictor` that has the largest R-square value?

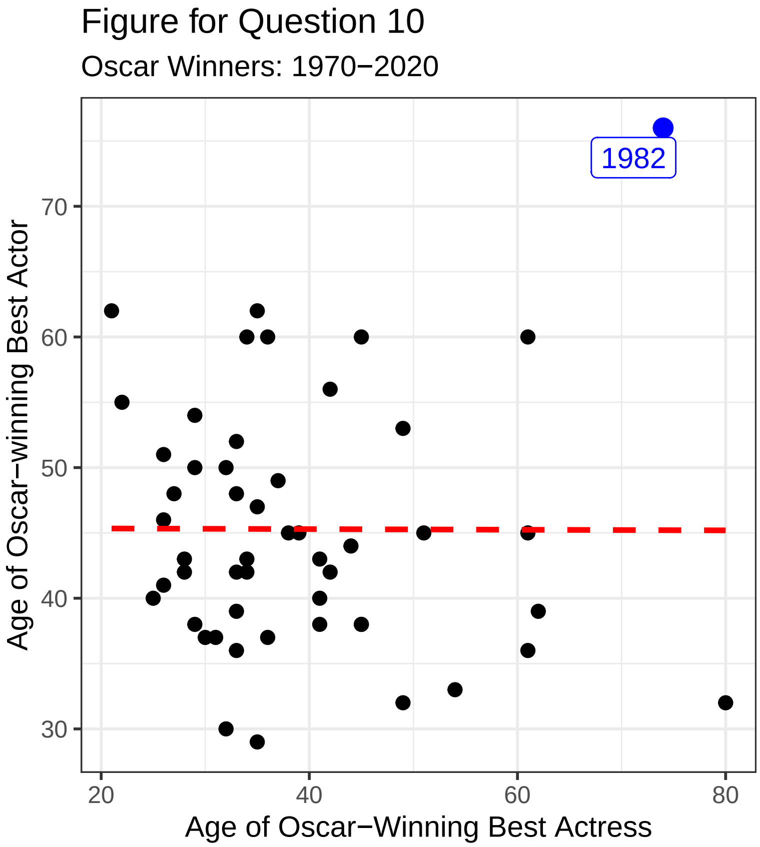
a. Plot A
b. Plot B
c. Plot C
d. Plot D

10 Question 10. (4 points)

The data in the `oscar.csv` file I have provided to you describe the winners of the Academy Awards (also called the "Oscars") for Best Actor and Best Actress from 1970 to 2020.

The Figure for Question 10 is a scatterplot of 51 points, in each case displaying the age of the Best Actor (on the vertical, or y, axis) and the age of the Best Actress (on the horizontal, or x, axis) from the Academy Awards. Note that the Pearson correlation coefficient associated with these data is -0.003.

```r
ggplot(oscar, aes(x = actress_age, y = actor_age)) +
    geom_point(size = 2) +
    geom_point(data = oscar %>% filter(year == 1982), col = "blue", size = 3) +
    geom_label_repel(data = oscar %>% filter(year == 1982),
                     aes(label = year), col = "blue") +
    geom_smooth(method = "lm", col = "red", lty = "dashed",
                se = FALSE, formula = y ~ x) +
    theme(aspect.ratio = 1) +
    labs(title = "Figure for Question 10",
         subtitle = "Oscar Winners: 1970-2020",
         x = "Age of Oscar-Winning Best Actress",
         y = "Age of Oscar-winning Best Actor")
```



Question 10 continues on the next page.

Continuation of Question 10

In 1982, Henry Fonda (age 76) and Katharine Hepburn (74) each won Oscars for the film *On Golden Pond*. This point is marked on the scatterplot by a blue dot, and labeled by its year. If the scatterplot were redrawn eliminating the 1982 awards, and including only the other 50 years, what would happen?

a. The slope of the linear model would DECREASE, and so would the model's R-squared.
b. The slope of the linear model would DECREASE, and the R-squared would INCREASE.
c. The slope of the linear model would INCREASE, and so would the R-squared.
d. The slope of the linear model would INCREASE, and the R-squared would DECREASE.
e. It is impossible to tell from the information provided.

11 Question 11.

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. Questions 11, 12 and 13 each are related to this study.

Among several variables measured in that study were the child's weight percentile (a value between 0 and 100, with heavier children having higher values) and the number of cigarettes smoked per day in the child's home, for 43 children, including 18 girls and 25 boys.

For the 18 girls in the study, the Pearson correlation coefficient between weight percentile and cigarettes smoked was reported as r = -0.50. Suppose that model A is a linear regression model predicting weight percentile using cigarettes smoked in the home for the 18 girls. Which of the following interpretations of this result is most correct?

a. The slope of a linear regression model predicting weight percentile using cigarettes smoked in the home for the 18 girls will be positive, and the resulting model will account for at least 50% of the variation in weight percentiles.
b. The slope of a linear regression model predicting weight percentile using cigarettes smoked in the home for the 18 girls will be negative, and the resulting model will account for at least 50% of the variation in weight percentiles.
c. The slope of a linear regression model predicting weight percentile using cigarettes smoked in the home for the 18 girls will be positive, and the resulting model will account for between 10% and 49% of the variation in weight percentiles.
d. The slope of a linear regression model predicting weight percentile using cigarettes smoked in the home for the 18 girls will be negative, and the resulting model will account for between 10% and 49% of the variation in weight percentiles.
e. The slope of a linear regression model predicting weight percentile using cigarettes smoked in the home for the 18 girls will be positive, and the resulting model will account for less than 10% of the variation in weight percentiles.
f. The slope of a linear regression model predicting weight percentile using cigarettes smoked in the home for the 18 girls will be negative, and the resulting model will account for less than 10% of the variation in weight percentiles.
g. None of these interpretations are correct.

12 Question 12. (4 points)

To help describe the 25 boys in the study we described in Question 11, I have provided the `tobacco_boys.csv` data file. I used those data to fit a least squares regression model to predict weight percentile (`weight.percentile` in the data set) on the basis of number of cigarettes smoked per day (`cigarettes`), summarized here.

```
q12 <- read_csv("data/tobacco_boys.csv")

q12 %>% head()
```

```
# A tibble: 6 x 3
     id weight_percentile cigarettes
  <dbl>             <dbl>      <dbl>
1     1                 6          0
2     2                 6         15
3     3                 2         40
4     4                 8         23
5     5                11         20
6     6                17          7
```

```
m1 <- lm(weight_percentile ~ cigarettes, data = q12)

tidy(m1, conf.int = TRUE, conf.level = 0.90) %>%
    select(term, estimate, conf.low, conf.high) %>% kable(digits = 3)
```

term	estimate	conf.low	conf.high
(Intercept)	41.153	29.425	52.881
cigarettes	-0.262	-0.896	0.373

```
glance(m1) %>% select(r.squared, sigma, AIC, nobs) %>%
    kable(digits = c(4, 1, 1, 0))
```

r.squared	sigma	AIC	nobs
0.0213	24.7	235.2	25

Question 12 continues on the next page.

Continuation of Question 12

Which of the following statements are true?

I. The model shows a mean prediction error of 24.7.

II. Each additional cigarette is associated with a 2.13% change in the variation of "weight_percentile".

III. Kids living where more cigarettes were smoked had larger weight percentile values, on average.

a. I only
b. II only
c. III only
d. I and II
e. I and III
f. II and III
g. I, II and III
h. None of these statements.

13 Question 13. (4 points)

Again, this question continues our work with the study described in Question 11. Information on the 25 boys in the study are provided to you in the `tobacco_boys.csv` data file. In Question 12, we used those data to fit a least squares regression model to predict weight percentile (`weight.percentile` in the data set) on the basis of number of cigarettes smoked per day (`cigarettes`).

Now, suppose a new child named Dan enters the study, and in his home, 24 cigarettes are smoked per day. Using the model `m1` fit in Question 12, specify the predicted (fitted) value for Dan, as an integer, in other words, rounded to zero decimal places.

14 Question 14. (4 points)

Suppose you have collected data as part of a cohort study to look at the impact of exposure to an industrial solvent (which is stored in a four-level character variable called `solvent` which can be either none, modest, moderate or profound) on the probability of a bladder cancer diagnosis (stored as a three-level character variable called `diagnosis` which can be either definite, possible, or no.)

You can assume that a tibble containing these variables called `q14` is available to you in R, and that the packages loaded by Dr. Love at the start of this Quiz are also already loaded for you, so you don't need to load them again and there should be no `library()` calls in your response. You may also assume that all of the R packages Dr. Love has asked you to install for this course are installed.

Provide a single line of R code (you may use at most two pipes) to obtain an appropriate numerical summary of the relationship between the `solvent` and `diagnosis` variables in the `q14` tibble.

15 Question 15. (4 points)

Suppose now that in continuing your work on the study from Question 14, you now have more granular information on the exposure level to the solvent. Specifically, you now have an `exposure` measure, expressed as the percentage of the Occupational Safety and Health Administration (OSHA) recommended exposure limit, so that 100 = the recommended exposure limit for this solvent, and values above 100 indicate exposures that exceed that limit, while values below 100 indicate exposures that are at least somewhat "safe".

You can assume that a new tibble, called `q15` is available to you in R containing this `exposure` measure as well as the bladder cancer `diagnosis` variable described in Question 14.

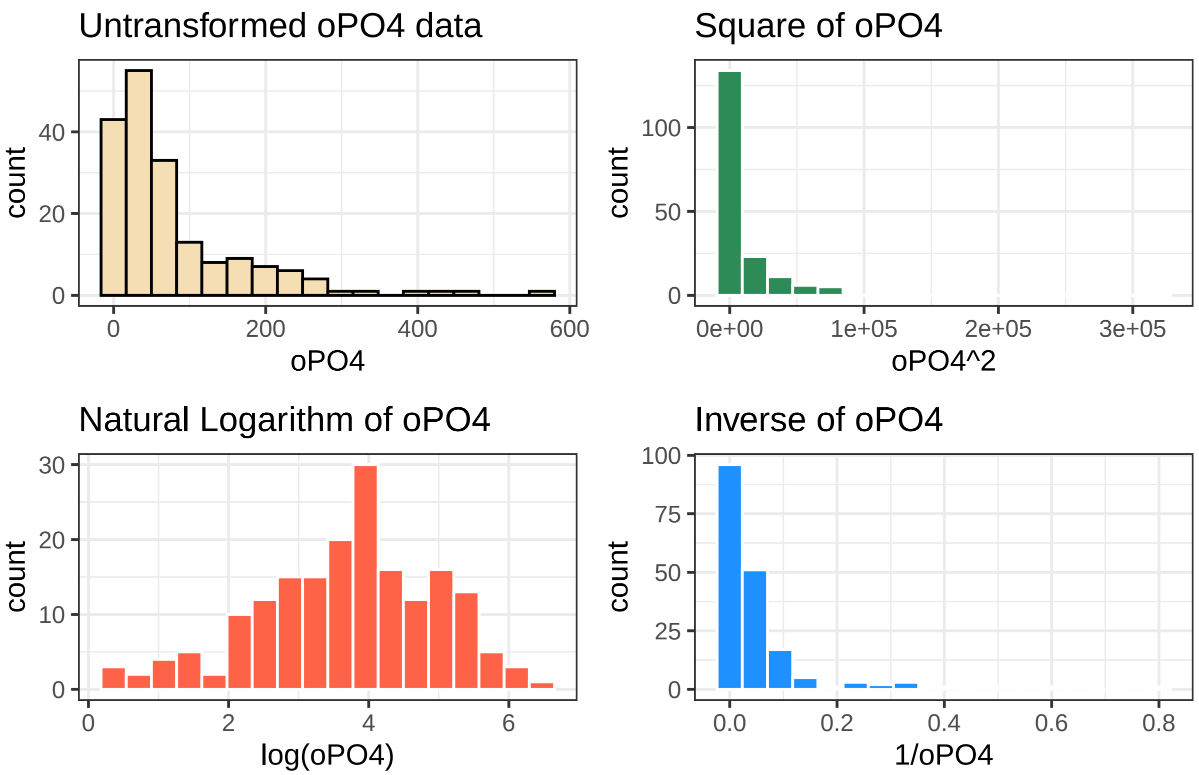
Again, you can also assume that the packages loaded by Dr. Love at the start of this Quiz are also already loaded for you, so you don't need to load them again. You may assume that all of the R packages Dr. Love has asked you to install for this course are installed, as well.

Provide a single line of R code (you may use at most two pipes) to obtain an appropriate numerical summary description of the distribution of `exposure` within each `diagnosis` group in the `q15` tibble.

16 Question 16.

Our next data set is called `algae.csv` and it is available to you with the Quiz materials. This data file describes 200 water samples collected from the same river over a period of 3 months, of which 184 have complete data. Each of those 184 observations contains information on a series of chemical parameters measured in the water samples, one of which is the mean value of orthophosphate, contained in the variable `oPO4`.

Figure for Question 16



Untransformed oPO4 data

Square of oPO4

Natural Logarithm of oPO4

Inverse of oPO4

Consider the histograms shown in the Figure for Question 16, and suppose your goal is to approximate a Normal distribution with some transformation of the `oPO4` data. Which of the following options describes the most logical transformation to use in trying to accomplish this goal?

a. The square of the oPO4 data
b. The natural logarithm of the oPO4 data
c. The inverse of the oPO4 data
d. The untransformed oPO4 data
e. It is impossible to tell from the information provided.

17 Question 17.

Return to the `algae.csv` file I provided to you, and fit a linear model to predict the natural logarithm of the algae frequency `a1` using the natural logarithm of the `oPO4` (orthophosphate) in the same water sample.

You will have to manage the data to use only those samples with complete data on both variables in your model, and which have values of `a1` that exceed zero.

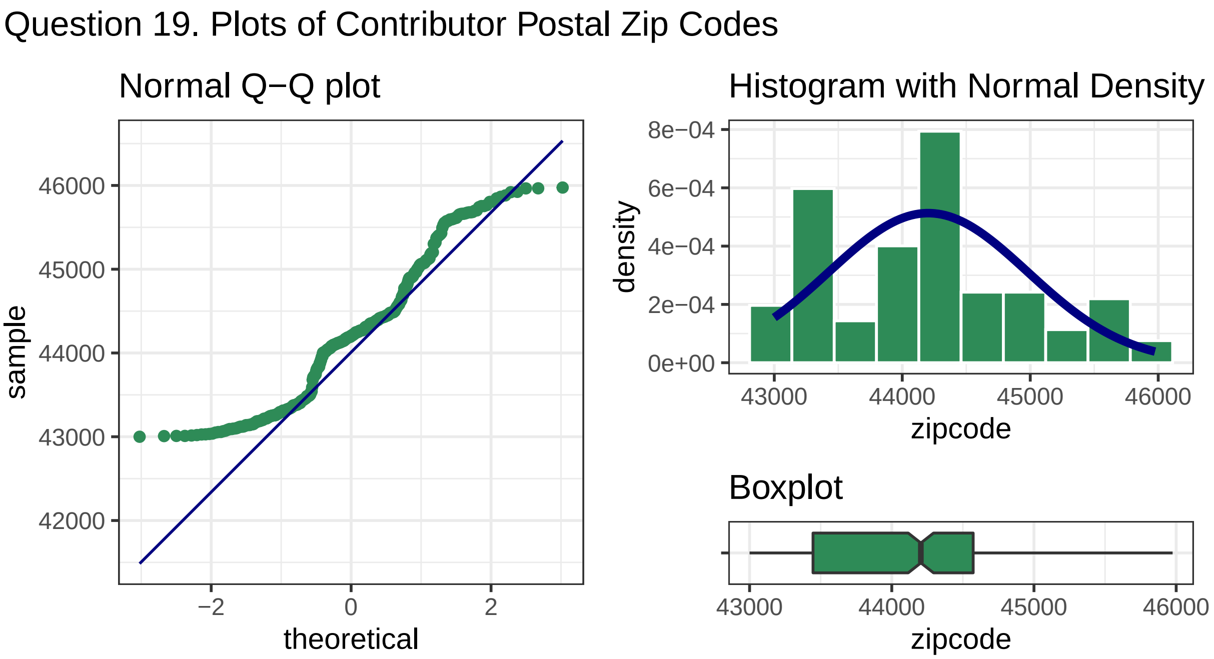
How many water samples are included in your model?

18 Question 18.

Based on your model described in Question 17, what is the predicted value of the actual algae frequency `a1` (be careful: what does your model predict?) for a sample with an `oPO4` value of 64? Round your response to a single decimal place.

19 Question 19.

The plot for Question 19 displays the postal zip codes of the last 400 Ohio residents who have made a disclosed individual financial contribution to a candidate in the 2020 presidential election.



Which of the following summaries of these data would be most appropriate?

a. The mean.
b. The median.
c. The interquartile range.
d. The mode.
e. It is impossible to tell.

20 Question 20. (4 points)

The table below shows the most recently measured body-mass index (BMI) of the 15 female patients that are scheduled to be seen this afternoon by a nurse practitioner for primary care of their chronic illness.

Patients scheduled for this afternoon

Patient	BMI (kg/mˆ2)	Height (m)	Weight (kg)
Allen, L	47.162534	1.65	128.4
Bieber, S	47.122586	1.63	125.2
Carrasco, C	38.220022	1.82	126.6
Civale, A	37.857802	1.71	110.7
Hand, B	34.726353	1.64	93.4
Hill, C	31.000918	1.65	84.4
Karinchak, J	30.884474	1.58	77.1
Maton, P	30.035003	1.63	79.8
McKenzie, T	29.703632	1.73	88.9
Perez, O	28.040197	1.63	74.5
Plesac, Z	27.952452	1.57	68.9
Plutko, A	27.813209	1.68	78.5
Quantrill, C	25.607639	1.44	53.1
Rodriguez, J	25.254996	1.63	67.1
Wittgren, N	20.974482	1.57	51.7
—	—	—	—
Average for these 15 Patients	32.534331	1.64	87.2
Practice Average across all Female Patients	31.169029	1.62	81.8

In one complete English sentence, suggest a worthwhile improvement to this table.

21 Question 21.

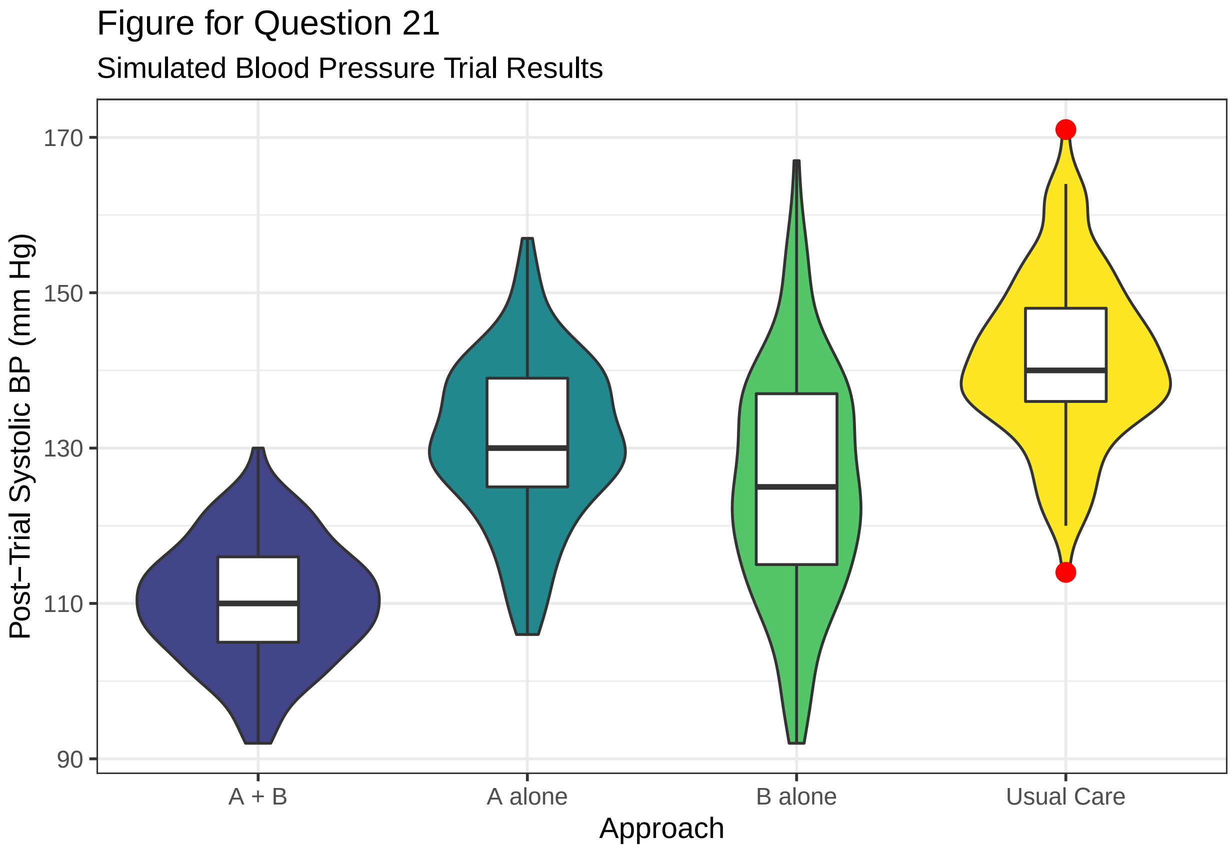
The initial results of the Systolic Blood Pressure Intervention Trial (SPRINT) received a lot of attention.

Quoting a press release from the National Institutes of Health:

> SPRINT evaluates the benefits of maintaining a new target for systolic blood pressure, the top number in a blood pressure reading, among a group of patients 50 years and older at increased risk for heart disease or who have kidney disease. A systolic pressure of 120 mm Hg, maintained by this more intensive blood pressure intervention, could ultimately help save lives among adults age 50 and older who have a combination of high blood pressure and at least one additional risk factor for heart disease, the investigators say.

Consider a hypothetical trial, where two different interventions are studied to see whether patients in another population besides that studied in SPRINT may have their blood pressure effectively managed to fall at the target level (120 mm Hg or lower).

500 patients were included in this trial, and were randomly allocated (125 to each intervention) so that we have 125 patients receiving both interventions A and B, 125 receiving A alone, 125 receiving B alone, and 125 receiving usual care (neither A nor B). The post-trial Systolic Blood Pressure results for all 500 patients are shown in the Figure for Question 21.



Question 21 continues on the next page.

Continuation of Question 21

Consider the following statements:

I. The group of patients receiving usual care had the smallest number of patients with SBP at 120 or lower after the trial.

II. The group of patients receiving B alone had the largest spread in their distribution of post-trial systolic blood pressures.

III. The group of patients receiving both A and B had more than 90 patients with post-trial SBP at 120 or lower.

Which of these statements are true?

a. I only
b. II only
c. III only
d. I and II
e. I and III
f. II and III
g. All three statements
h. None of the three statements

22 Question 22.

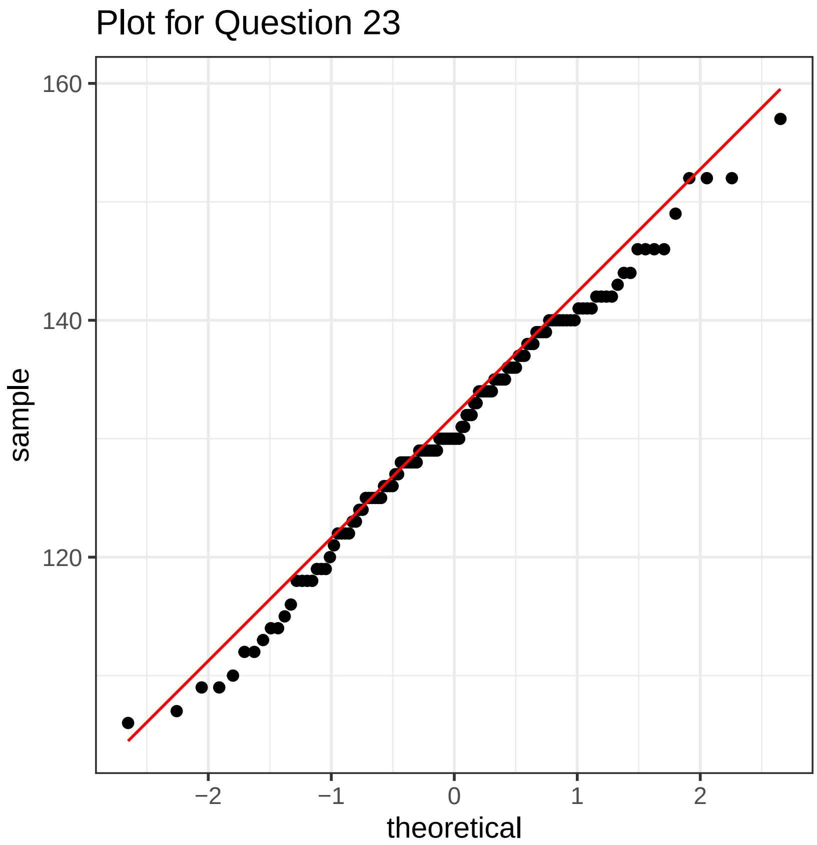
Which of the four blood pressure trial groups discussed in Question 21 produced the individual subject with the lowest post-trial systolic blood pressure?

a. The group receiving A alone
b. The group receiving B alone
c. The group receiving usual care
d. The group receiving both A and B
e. It is impossible to tell from the information provided.

23 Question 23.

The normal Q-Q plot shown here is taken from one of the four blood pressure trial groups discussed in Questions 21 and 22. Which one?

```
bp_trial %>% filter(approach == "A alone") %>%
    ggplot(., aes(sample = sbp_post)) +
    geom_qq() + geom_qq_line(col = "red") +
    theme(aspect.ratio = 1) +
    labs(title = "Plot for Question 23")
```



a. The group receiving A alone
b. The group receiving B alone
c. The group receiving usual care
d. The group receiving both A and B

24 Question 24. (4 points)

Consider the `starwars` tibble that is part of the `dplyr` package in the tidyverse. Use those data to address Question 24 and Question 25.

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 humans with blue `eye_color` and brown `hair_color`, specifically, here, and not with other related colors or combinations of these with other colors.)

25 Question 25. (4 points)

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

26 Question 26. (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 26 (found on the next page of this Quiz) does not include a problem of this sort.

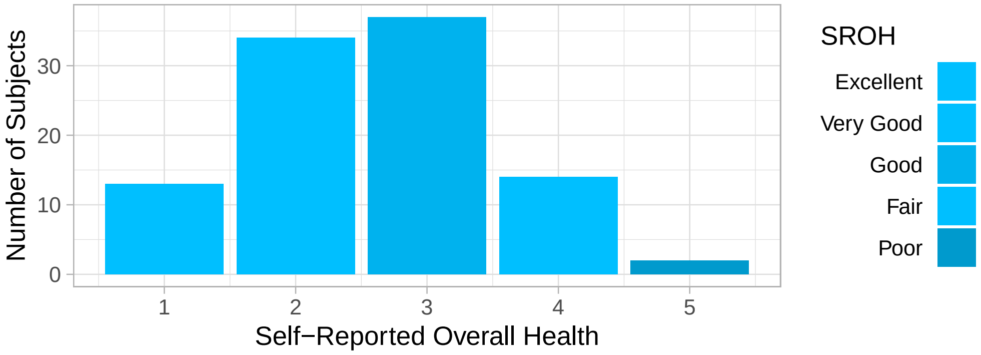
Please identify the **good** plot - the one that avoids Jeff's pitfalls.

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

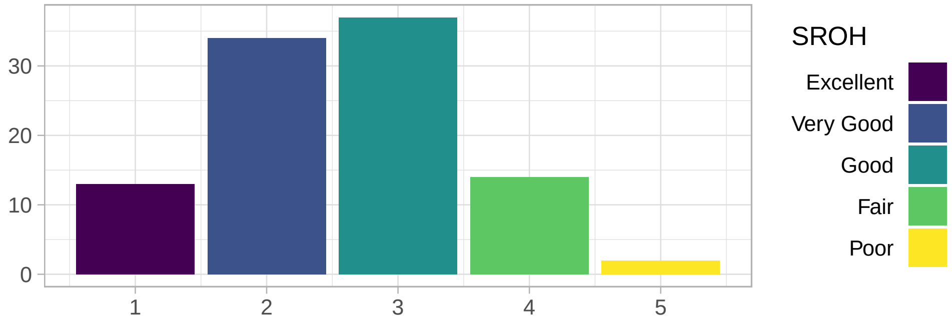
Again, the Figure for Question 26 is shown on the next page.

Figure for Question 26

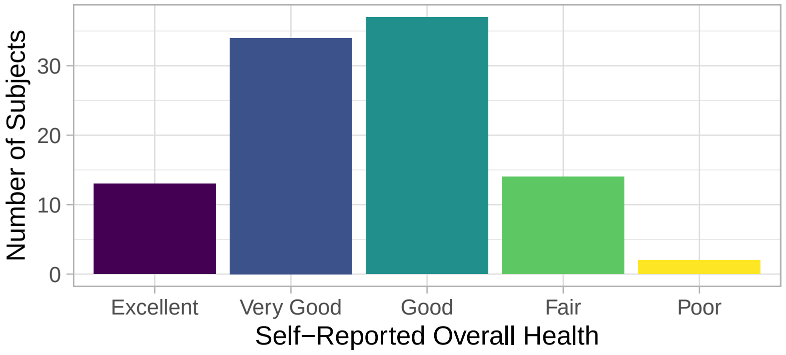
Plot A. Few Subjects report Poor Health



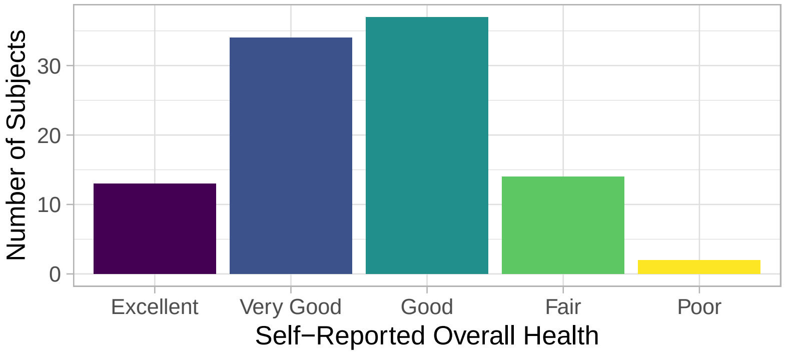
Plot B. Bar Chart for Question 26



Plot C. Few Subjects report Poor Health



Plot D. Bar Chart for Question 26

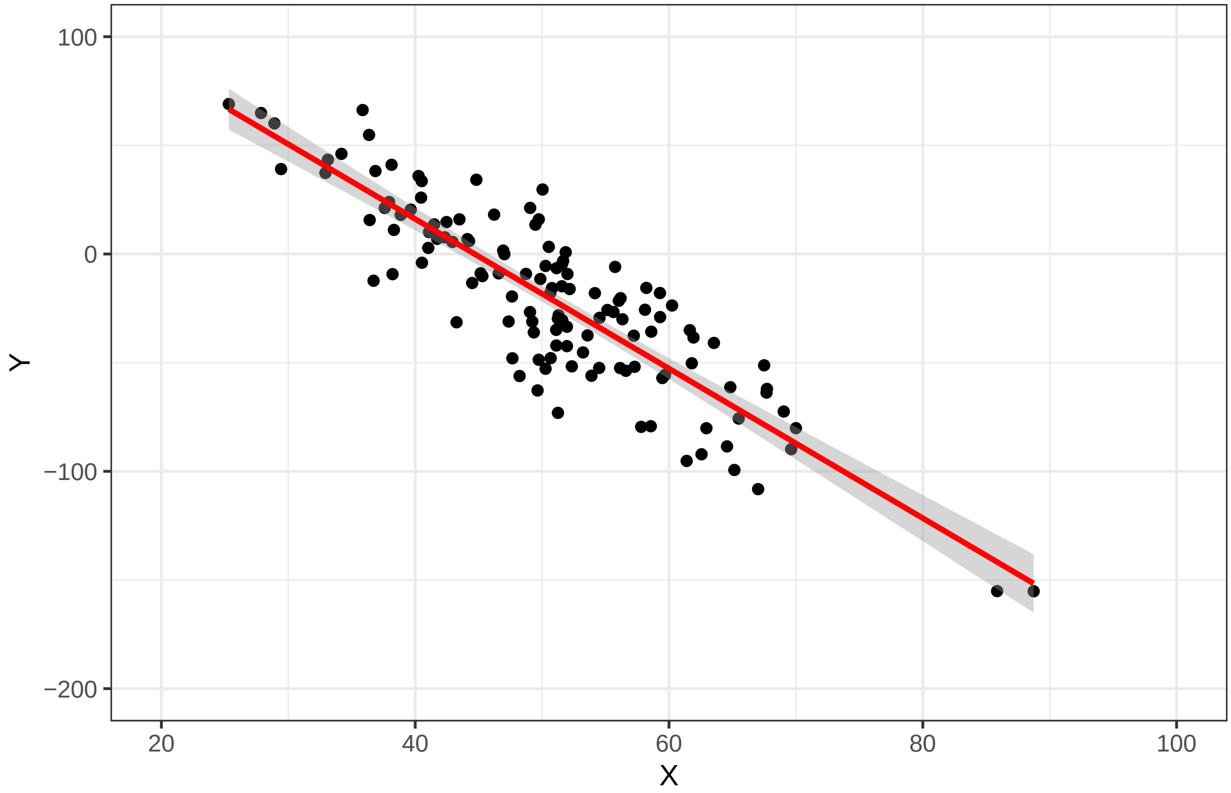


27 Question 27.

Consider the following possible summaries of a linear model fit to predict Y from X, describing the scatterplot shown in the Figure for Question 27. Which of these summaries is correct?

a. Model: y = 3.4 + 154 x, with R-squared = -0.76
b. Model: y = 3.4 - 154 x, with R-squared = -0.26
c. Model: y = -3.4 + 154 x, with R-squared = 0.76
d. Model: y = -3.4 + 154 x, with R-squared = 0.26
e. Model: y = 3.4 + 154 x, with R-squared = 0.76
f. Model: y = 3.4 + 154 x, with R-squared = 0.26
g. Model: y = 154 - 3.4 x, with R-squared = -0.76
h. Model: y = 154 - 3.4 x, with R-squared = -0.26
i. Model: y = 154 + 3.4 x, with R-squared = 0.76
j. Model: y = 154 + 3.4 x, with R-squared = 0.26
k. Model: y = 154 - 3.4 x, with R-squared = 0.76
l. Model: y = 154 - 3.4 x, with R-squared = 0.26

Figure for Question 27



28 Question 28.

According to the *Elements of Data Analytic Style*, which of the following elements belong in a proper data analysis report? (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.

29 Question 29. (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 29 describes the fat (in g) and sodium (in mg) contents of twelve 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 for Question 29?

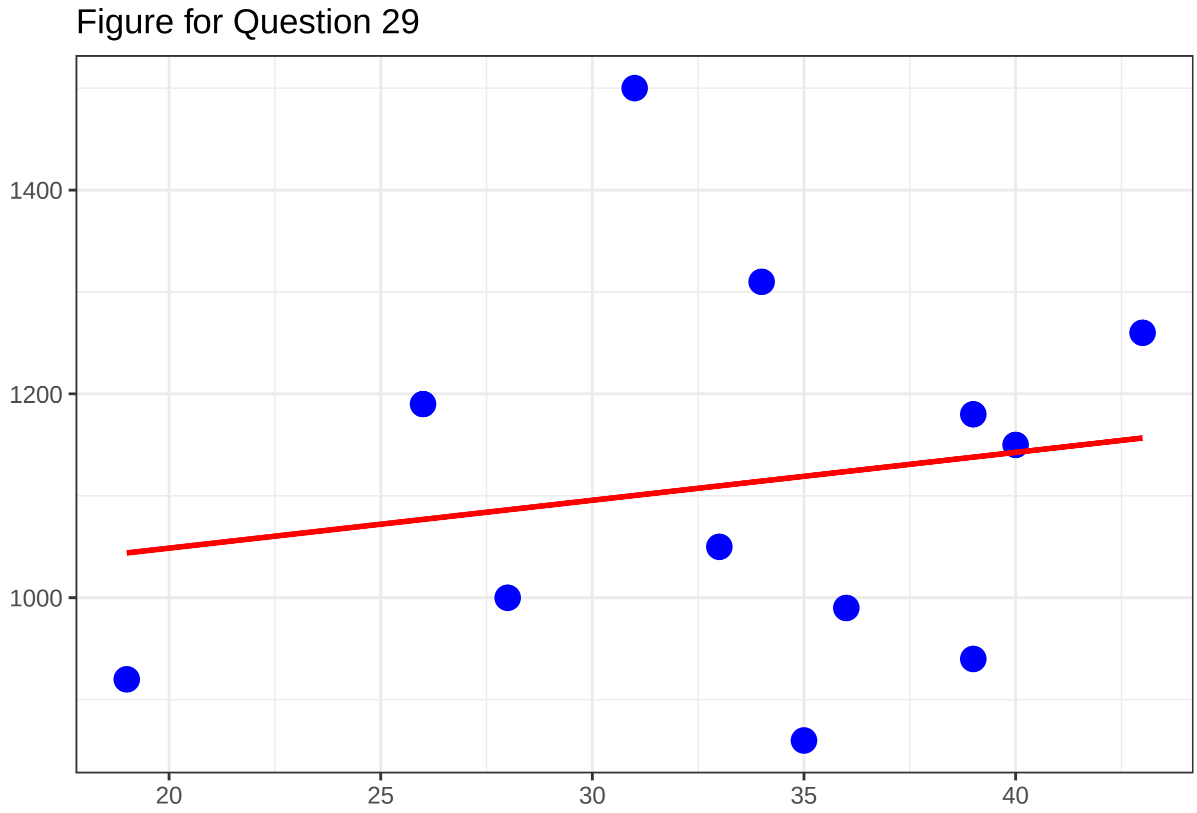


Figure for Question 29