431: Quiz 1 Sketch and Results

Thomas E. Love

Created 2020-10-07 20:48:21

Contents

R Packages Dr. Love Used To Build The Sketch

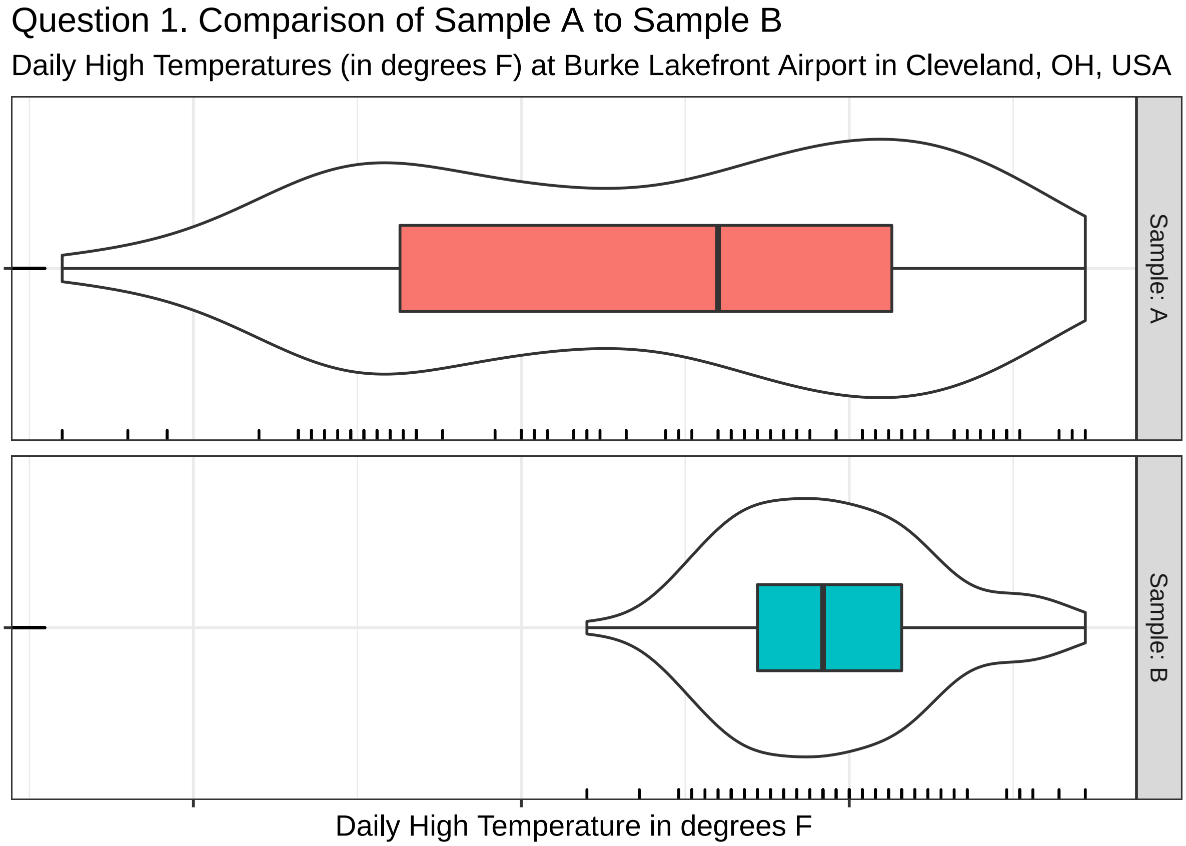
```r
library(broom)
library(car)
library(ggrepel)
library(janitor)
library(knitr)
library(magrittr)
library(patchwork)
library(tidyverse)

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

1.1 Answer 1 is f.

Statement I is false. September is a relatively warm month in Cleveland (much warmer, for instance, than the winter months.) Of these two samples, Sample B is clearly centered at a much higher temperature.

Statement II is true. We can see from the boxplots that the width of the box (which is the IQR) in Sample A is clearly greater than that shown for Sample B.

Statement III is also true. The Sample A data do not appear to describe a Normal distribution. There is, for instance, no central peak to that distribution as we can see from the violin plot. The Sample A data appear more uniform than a Normal or perhaps might be interpreted as multi-modal. Sample B, on the other hand, is reasonably symmetric, has a central peak, and no obvious signs of unusual tail behavior.

1.1.1 Grading 1

Question	Value	% Correct	% Points
1	4 points	87.1	90.7

- Most common incorrect response was *b*.
- Also selected: `e`.

For each question, I provide a table like the one shown above, containing the Question number, its value (maximum score is either 3 points or 4 points), the % of students who received full credit (% Correct), and the % of total available points that were awarded (including, where available, partial credit.)

Next, for multiple choice questions only, I show the most common incorrect response, and the other responses that were selected.

- I gave 2 points of partial credit for response `b` here.

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`.

2.1 Answer 2 is e.

Every extra square foot of Area is associated with an additional $35.30 in average price, for houses with a given Lotsize and Age. This is what the model tells us, based on the slope of Area, after adjusting for Lotsize and Age. So d is correct. The other responses are not. In particular, `b` is not appropriate because it doesn't require us to hold the other predictors constant when we make our comparison.

2.1.1 Grading 2

Question	Value	% Correct	% Points
2	3 points	72.9	72.9

- Most common incorrect response was *b*.
- Also selected: `d`, `a`.

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

3.1 Answer 3 is that I is a, II is b, and III is a

See Spiegelhalter, Chapter 3.

Statement I describes the principle of validity, in the sense (as Spiegelhalter puts it) of measuring what you really want to measure and not having a systematic bias. This is part of moving from raw data to the sample, so I is primarily associated with a.

Statement II is about the importance of random sampling in order to generate a representation of the study population from the data, so II is primarily associated with b.

Statement III describes the principle of reliability, in the sense (as Spiegelhalter puts it) of having low variability from occasion to occasion, and so being a precise or repeatable number. This is again part of moving from raw data to the sample, so III is also primarily associated with a.

3.1.1 Grading 3

Question	Value	% Correct	% Points
3 (I)	1 points	88.6	88.6
3 (II)	1 point	87.1	87.1
3 (III)	1 point	22.9	22.9

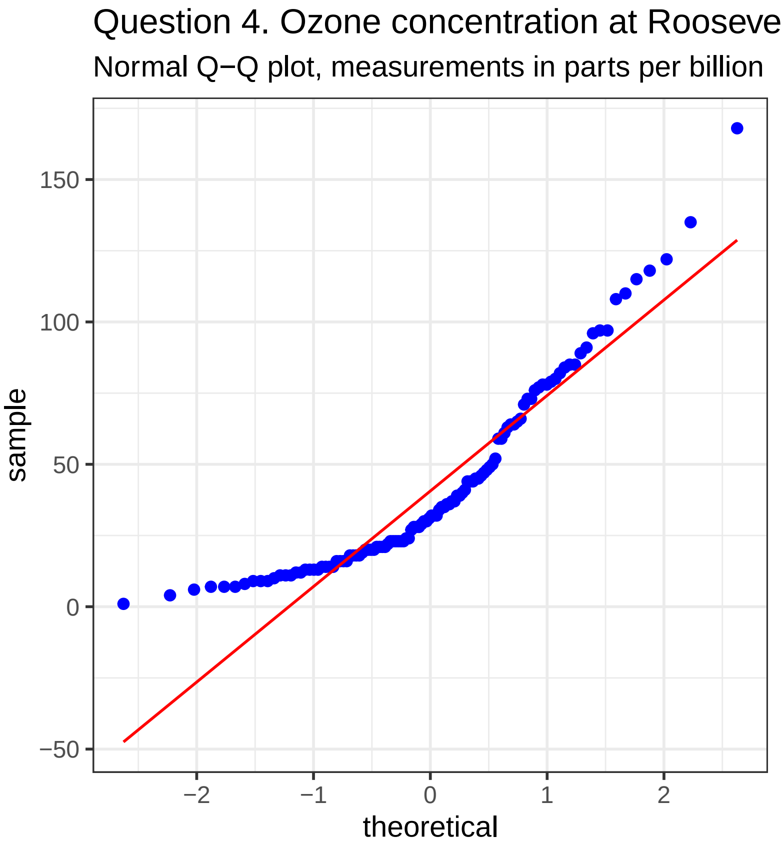
- For part (I), the incorrect response people chose was `b`
- For part (II), `a`, then `c` were both mentioned
- For part (III), almost everyone who was incorrect picked `c`.

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.



Question 4. Ozone concentration at Roosevelt Island

Normal Q−Q plot, measurements in parts per billion

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.

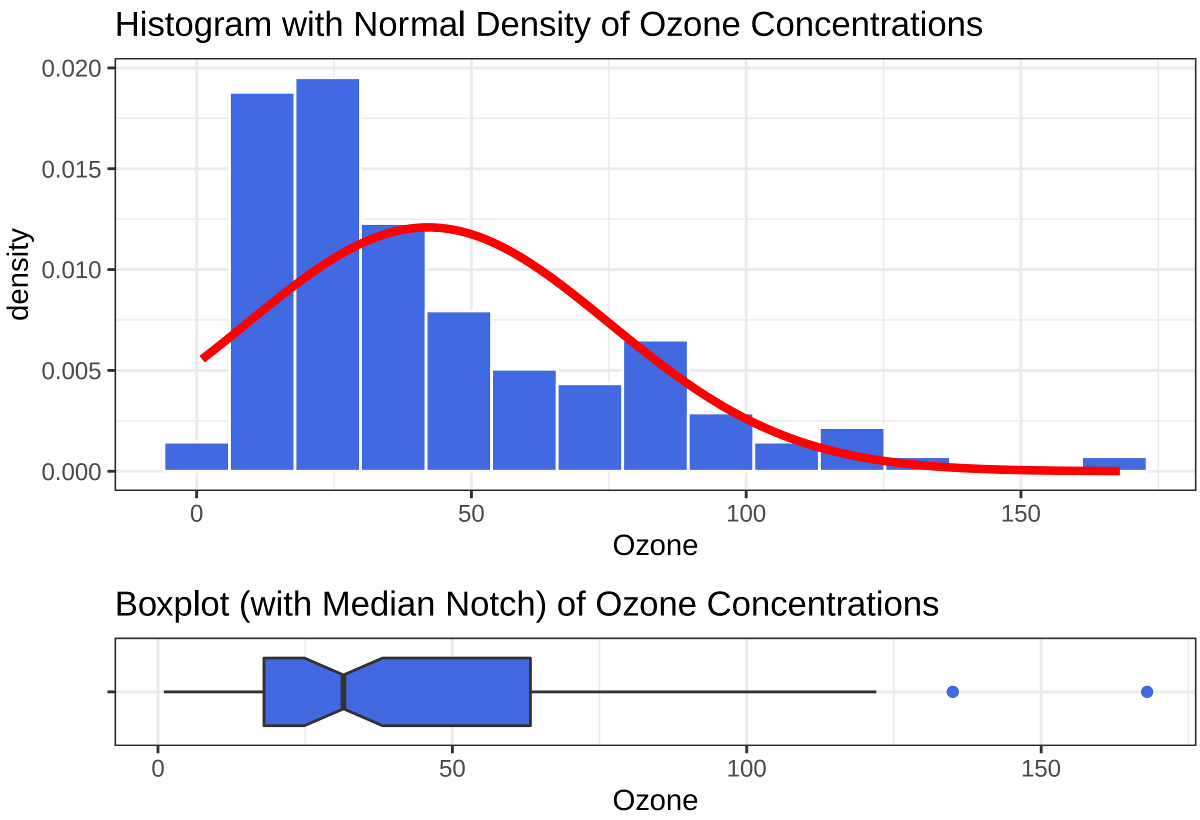
4.1 Answer 4 is c.

The data are right skewed, by the definition of how a Normal Q-Q plot works. This is perhaps easiest to demonstrate by plotting the same data in other ways.

```
p1 <- ggplot(ozone, aes(x = Ozone)) +
  geom_histogram(aes(y = stat(density)),
                 bins = 15, fill = "royalblue", col = "white") +
  stat_function(fun = dnorm,
                args = list(mean = mean(ozone$Ozone),
                            sd = sd(ozone$Ozone)),
                col = "red", lwd = 1.5) +
  labs(title = "Histogram with Normal Density of Ozone Concentrations")

p2 <- ggplot(ozone, aes(x = Ozone, y = "")) +
  geom_boxplot(fill = "royalblue", outlier.color = "royalblue", notch = TRUE) +
  labs(title = "Boxplot (with Median Notch) of Ozone Concentrations", y = "")

p1 / p2 + plot_layout(heights = c(4,1))
```



4.1.1 Grading 4

Question	Value	% Correct	% Points
4	4 points	94.3	94.3

- Most common incorrect response was *e*.
- Also selected: `d`.

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.

5.1 Answer 5 is a.

The mean is located at the value of the sample that corresponds to the theoretical quantile at 0 from the standard Normal distribution (which has mean 0 and standard deviation 1). That value is clearly below 50, so the mean is below 50 ppb. In fact, the actual mean ozone concentration is 42.1 ppb.

The IQR is the middle half of the data, which we don't see directly in the Normal Q-Q plot. However, we can look at the range from -1 to +1 in the theoretical quantiles, which corresponds to a larger distance, because it covers the middle 68% of the data. At x = -1, the concentration is about 20 and at x = +1, it's about 70, so the IQR must be smaller than the difference (70-20), and so the IQR is below 50 ppb, let alone 75 ppb. In fact, the IQR for ozone concentration turns out to be 45.25, based on a 25th percentile of 18 and a 75th percentile of 63.25.

Source The data used in Questions 5-6 come from an example in van Belle G Fisher L Heagerty PJ and Lumley T *Biostatistics: A Methodology for the Health Sciences* (2nd Edition), Chapter 3.

5.1.1 Grading 5

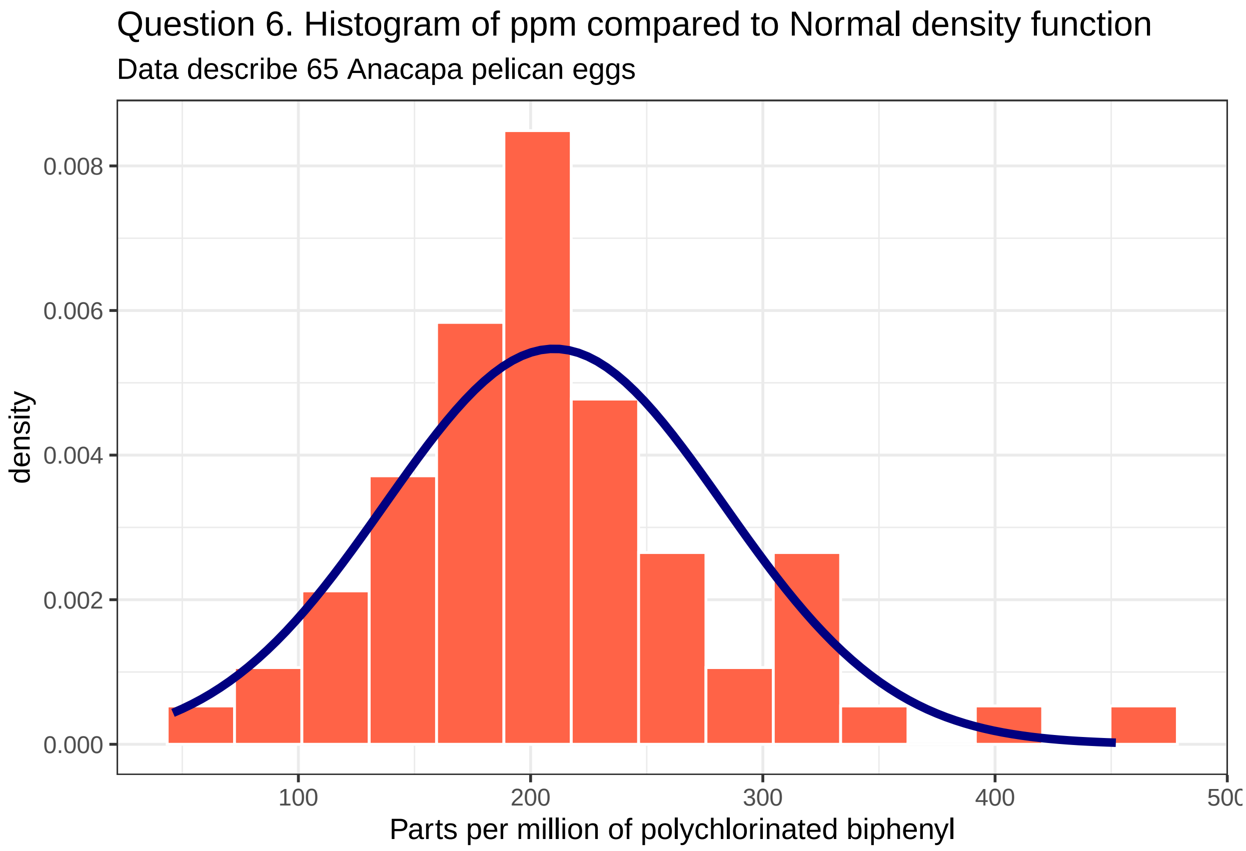
Question	Value	% Correct	% Points
5	3 points	44.3	44.3

- Most common incorrect response was *e*.
- Also selected: `b` and `c`.

This was one of the three questions I thought would be "trickiest" prior to you seeing the Quiz, and it definitely was difficult. I assume this is because it's requiring you to use the Normal Q-Q plot in a different way than we might usually.

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")
```

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

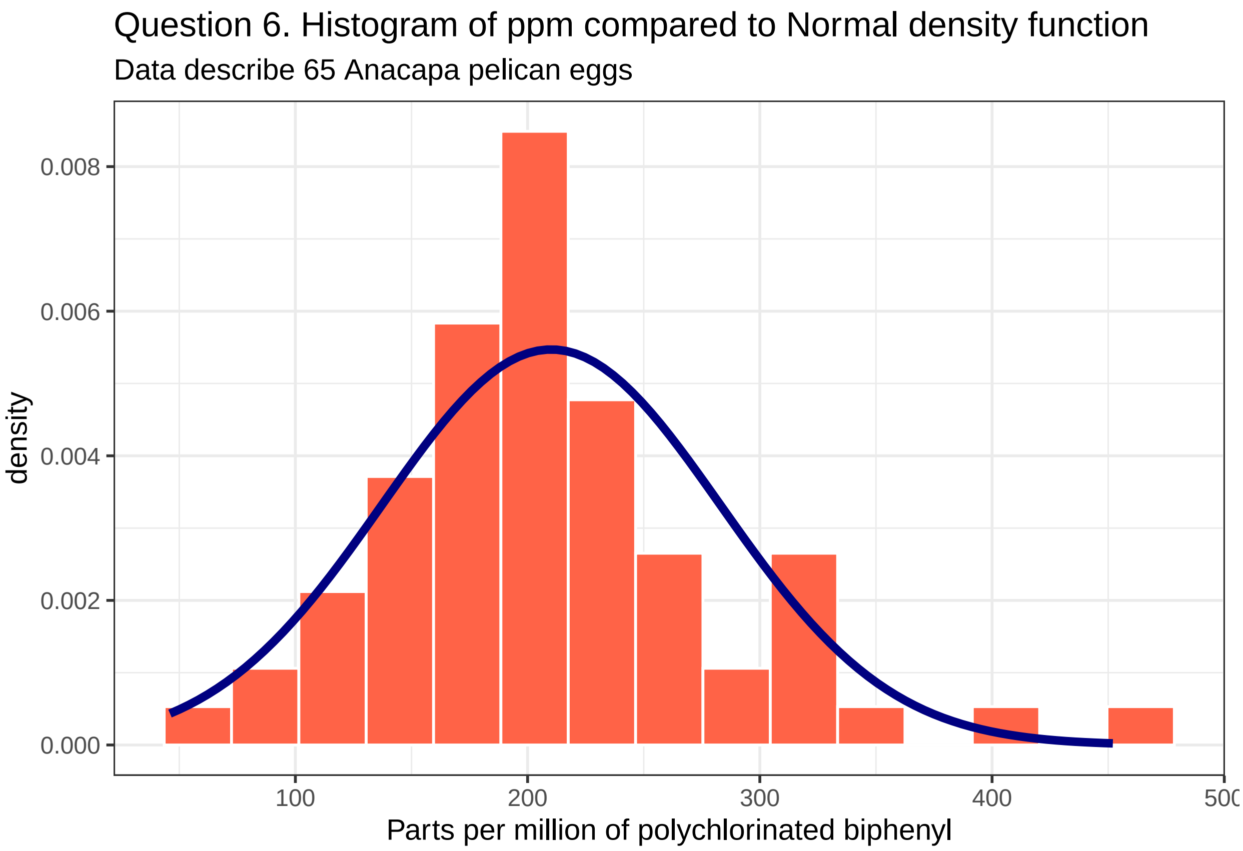
b. Line 4
c. Line 5
d. Line 6
e. Line 7

6.1 Answer 6 is that a c and e should be dropped.

Here's the code that generated the plot. As you can see, the hashtag comments out what were described as lines 3, 5 and 7. Including any of those lines, or dropping any of the others, changes the plot in clear ways.

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

ggplot(pelican, aes(x = ppm)) +
#    geom_density(col = "navy", lwd = 1.5) +
    geom_histogram(aes(y = stat(density)), bins = 15, fill = "tomato", col = "white") +
#    geom_histogram(bins = 15, fill = "tomato", col = "white") +
    stat_function(fun = dnorm,
                  args = list(mean = mean(pelican$ppm), sd = sd(pelican$ppm)),
                  col = "navy", lwd = 1.5) +
#    coord_flip() +
    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")
```



6.1.1 Grading 6

Question	Value	% Correct	% Points
6	4 points	64.3	81.8

In this question, you had 5 decisions to make (include or exclude lines 3, 4, 5, 6 and 7.) I awarded…

- 4 points if you made all 5 decisions correctly.
- 3 points for making 4 out of 5 correctly.
- 2 points for making 3 out of 5 correctly.
- 1 point for making 2 out of 5 correctly.
- Otherwise, zero points.

The most common incorrect responses were

- `a`, `b` and `e`, which was worth 2 points, and
- `c`, `d` and `e`, which was also worth 2 points

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.

7.1 Answer 7 is a.

The outcome goes on the vertical (Y) axis, and the predictor goes on the X axis. Here, we're modeling PCB concentration (our outcome) as a function of the shell thickness (our predictor). If we're just finding a correlation, it wouldn't matter, but if we're fitting a regression or other model, it does matter.

Data Source for Questions 6-7: Data set 165 in Hand DJ et al *A Handbook of Small Data Sets, Volume 1*. From Risebrough RW (1972) Effects of environmental pollutants upon animals other than man. *Proceedings of the 6th Berkeley Symposium on Mathematics and Statistics, VI.* California: University of California Press, 443-463.

7.1.1 Grading 7

Question	Value	% Correct	% Points
7	3 points	77.1	77.1

- Most common incorrect response was *b*.
- Also selected: `e` and `d`.

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

8.1 Answer 8 is b.

Both the mean and median would increase, in this case. Only the standard deviation would decrease. Here's the demonstration.

```
msage <- tibble(age_onset = c(14, 16, 17, 18, 18, 18, 18, 19,
                              20, 23, 26, 27, 32, 33, 39, 42, 44))

mosaic::favstats(~ age_onset, data = msage)
```

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

```
 min Q1 median Q3 max     mean       sd  n missing
  14 18     20 32  44 24.94118 9.653177 17       0
```

```
msage_2 <- msage %>% add_row(age_onset = 28)

mosaic::favstats(~ age_onset, data = msage_2)
```

```
 min Q1 median Q3 max     mean       sd  n missing
  14 18   21.5 31  44 25.11111 9.392669 18       0
```

Data Source: Data set 198 in Hand DJ et al *A Handbook of Small Data Sets, Volume 1*. From Joseph L et al (1990) Is multiple sclerosis an infectious disease? Inference about an input process based on the output. *Biometrics*, **46**, 337-349.

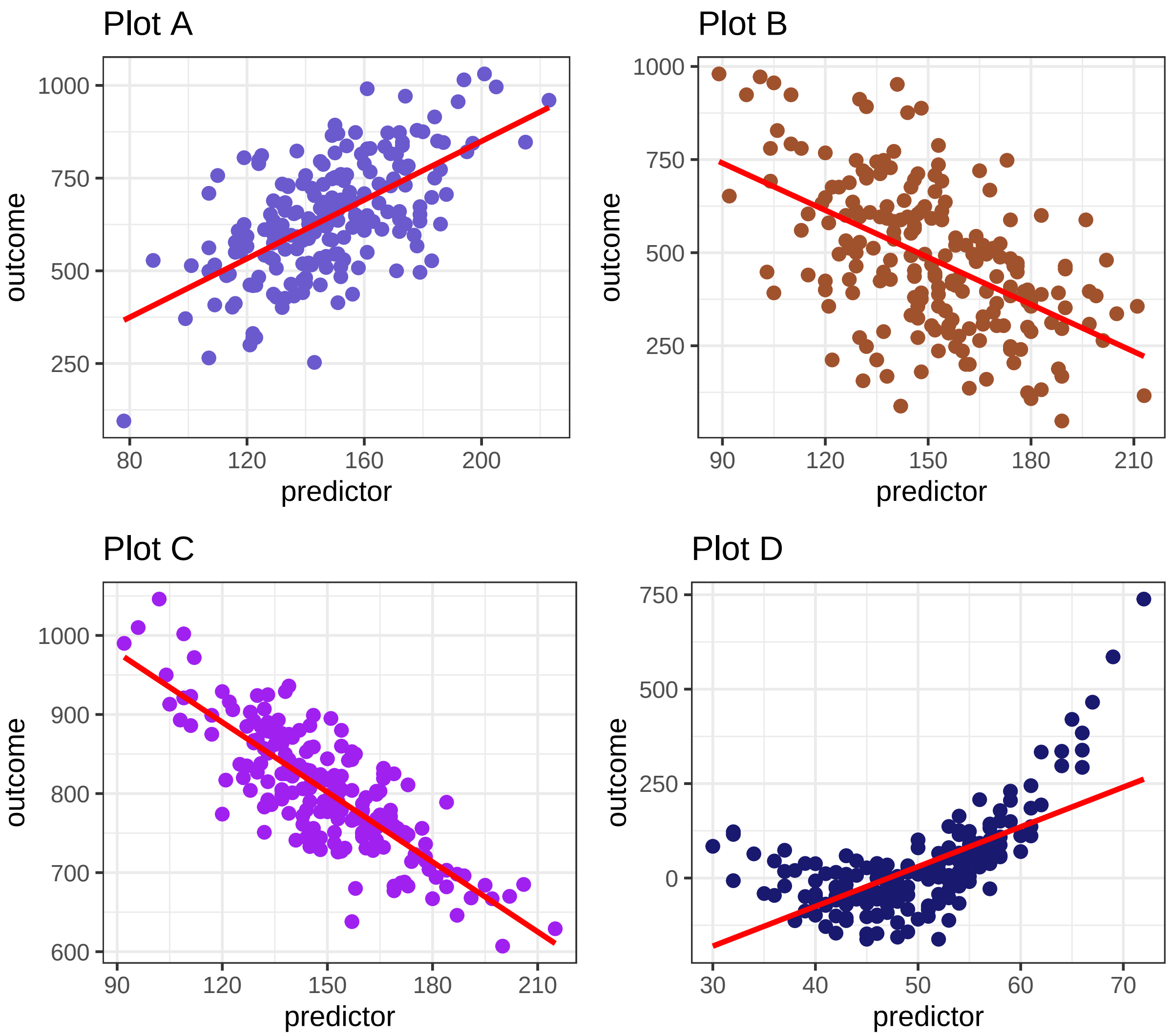
8.1.1 Grading 8

Question	Value	% Correct	% Points
8	3 points	85.7	85.7

- Most common incorrect response was *h*.
- Also selected: `d` and `g`.

9 Question 9.

Consider the four scatterplots provided for Question 9.



Plots for Question 9

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

9.1 Answer 9 is c.

Plots A and B show considerably weaker linear associations than Plot C. Plot D shows a fairly clearly non-linear association. The correlations and resulting R^2 values for the data in each plot are tabulated below.

Plot	Pearson Correlation	R-Square
A	0.604	0.365
B	-0.525	0.275
C	-0.815	0.664
D	0.646	0.418

9.1.1 Grading 9

Question	Value	% Correct	% Points
9	3 points	≥ 95	≥ 95

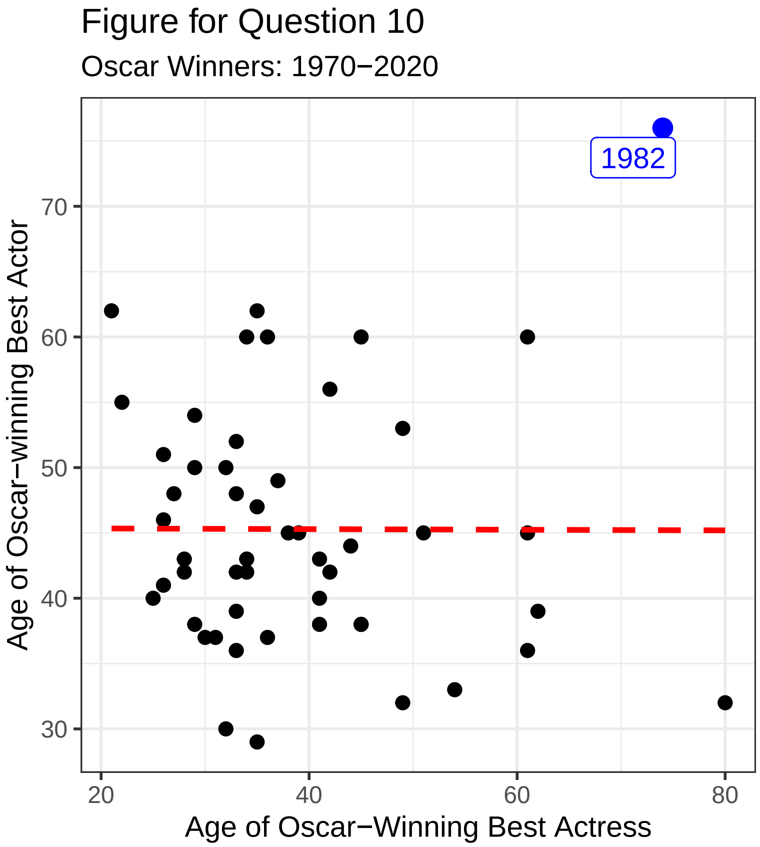
- Most common incorrect response was *a*.
- Also selected: `b`.

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.

```
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")
```



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.

10.1 Answer 10 is b.

Dropping the 1982 outlier will have a substantial effect on the regression line. When applied to the revised data, a regression line will fit the remaining 50 years substantially better (so the R^2, the square of the Pearson correlation, will increase) and the line will rise on the left of the graph and fall on the right, (so that the slope will be substantially decreased).

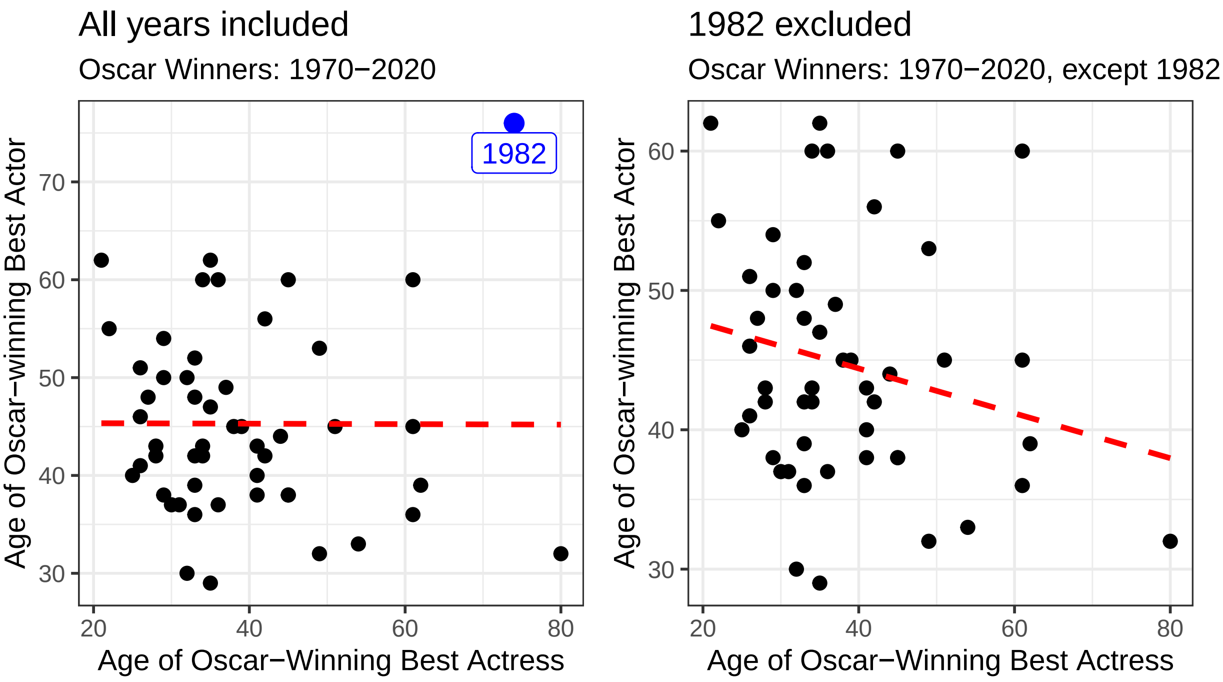
Here's what the plot looks like with and without 1982.

```r
p1 <- 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 = "All years included",
         subtitle = "Oscar Winners: 1970-2020",
         x = "Age of Oscar-Winning Best Actress",
         y = "Age of Oscar-winning Best Actor")

oscar_drop82 <- oscar %>% filter(year != 1982)

p2 <- ggplot(oscar_drop82, aes(x = actress_age, y = actor_age)) +
    geom_point(size = 2) +
    geom_smooth(method = "lm", col = "red", lty = "dashed",
                se = FALSE, formula = y ~ x) +
    theme(aspect.ratio = 1) +
    labs(title = "1982 excluded",
         subtitle = "Oscar Winners: 1970-2020, except 1982",
         x = "Age of Oscar-Winning Best Actress",
         y = "Age of Oscar-winning Best Actor")

p1 + p2
```



Here are the models, first including 1982…

```
m1 <- lm(actor_age ~ actress_age, data = oscar)

glance(m1) %>% select(r.squared, nobs)
```

```
# A tibble: 1 x 2
   r.squared  nobs
       <dbl> <int>
1 0.00000984    51
```

```
tidy(m1, conf.int = TRUE, conf.level = 0.90) %>%
    select(term, estimate, std.error, conf.low, conf.high) %>%
    kable(digits = 4)
```

term	estimate	std.error	conf.low	conf.high
(Intercept)	45.3865	4.4218	37.9731	52.7998
actress_age	-0.0024	0.1079	-0.1833	0.1786

and now without 1982…

```
m2 <- lm(actor_age ~ actress_age, data = oscar_drop82)

glance(m2) %>% select(r.squared, nobs)
```

```
# A tibble: 1 x 2
  r.squared  nobs
```

```
      <dbl> <int>
1    0.0484    50
```

```
tidy(m2, conf.int = TRUE, conf.level = 0.90) %>%
    select(term, estimate, std.error, conf.low, conf.high) %>%
    kable(digits = 4)
```

term	estimate	std.error	conf.low	conf.high
(Intercept)	50.8392	4.1235	43.9233	57.7552
actress_age	-0.1610	0.1030	-0.3337	0.0118

Summarizing in a table, we have:

Dates Included	Model for Actor Age	R^2 (as a %)
1970-2020 (with 1982)	45.4 - 0.002 Actress Age	< 0.001
1970-1981, 1983-2020	50.8 - 0.161 Actress Age	4.8

10.1.1 Grading 10

Question	Value	% Correct	% Points
10	3 points	81.4	81.4

- Most common incorrect response was *c*.
- Also selected: `d` and `a`.

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.

11.1 Answer 11 is d.

The slope of a 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.

This is because the Pearson correlation and the slope have the same sign, and because squaring the Pearson correlation gives R2 = (-0.5)(-0.5) = 0.25 so that the model will account for 25% of the variation in weight percentiles, and 25% is in fact between 10% and 49%.

11.1.1 Grading 11

Question	Value	% Correct	% Points
11	3 points	70.0	70.0

- Most common incorrect response was *g*.

- Also selected: `b`.

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

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.

12.1 Answer 12 is h.

None of these statements are true.

- Statement I is false. The σ value does not describe the average prediction error. It describes the residual standard error. With least squares (as `lm` uses to fit a model) the mean prediction error (i.e. the mean of the prediction errors, and nothing more) will always be 0.
- Statement II is, well, nonsense. The "variation of weight percentile" doesn't actually mean anything.
- Statement III is false. Actually, the slope is negative, so the opposite is true.

12.1.1 Grading 12

Question	Value	% Correct	% Points
12	3 points	48.6	48.6

- Most common incorrect response was *a*.
- Also selected: `b`, `d`, `f` and `h`.

People are always afraid to select "None of the above" it seems.

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.

13.1 Answer 13 is 35.

```
augment(m1, newdata = tibble(cigarettes = 24))
```

```
# A tibble: 1 x 2
  cigarettes .fitted
       <dbl>   <dbl>
1         24    34.9
```

and we can round this prediction from 34.87 up to 35, as requested.

13.1.1 Grading 13

Question	Value	% Correct	% Points
13	4 points	≥ 95	≥ 95

- Failing to round cost you a point.

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.

14.1 Answer 14 is a line of R code

We could create a little table to be sure.

```
q14 <- tibble(solvent = c("none", "none", "modest", "moderate", "moderate",
                          "moderate", "moderate", "profound", "profound",
                          "profound"),
              diagnosis = c("definite", "possible", "no", "no", "possible",
                            "possible", "definite", "possible", "definite",
                            "definite"))
```

I had in mind

```
q14 %>% tabyl(solvent, diagnosis)
```

```
  solvent definite no possible
 moderate        1  1        2
   modest        0  1        0
     none        1  0        1
 profound        2  0        1
```

perhaps adding the marginal totals with

```
q14 %>% tabyl(solvent, diagnosis) %>%
  adorn_totals(where = c("row", "col"))
```

```
  solvent definite no possible Total
 moderate        1  1        2     4
   modest        0  1        0     1
     none        1  0        1     2
 profound        2  0        1     3
    Total        4  2        4    10
```

or perhaps

```
table(q14$solvent, q14$diagnosis)
```

```
          definite no possible
  moderate        1  1        2
  modest          0  1        0
  none            1  0        1
  profound        2  0        1
```

or maybe

```
xtabs(~ solvent + diagnosis, data=q14)
```

```
          diagnosis
solvent    definite no possible
  moderate        1  1        2
  modest          0  1        0
  none            1  0        1
  profound        2  0        1
```

to produce a table with the solvent information in the rows and the diagnosis information in the columns. Inverting the order in the function calls above would produce the same table but switch the rows and columns, which is also fine.

Some people wanted to reorder the levels of the variables. I agree that might be valuable, but that would require something that really stretches the definition of a single line of code further than I had intended, although it can be done with just two pipes.

```
q14 %>% mutate(
  solvent_f = fct_relevel(factor(solvent), "none", "modest",
                          "moderate", "profound"),
  diagnosis_f = fct_relevel(factor(diagnosis),
                            "definite", "possible", "no")) %>%
  tabyl(solvent_f, diagnosis_f)
```

```
 solvent_f definite possible no
      none        1        1  0
    modest        0        0  1
  moderate        1        2  1
  profound        2        1  0
```

So if you succeeded in doing that, great. But it was not necessary.

Some other people wanted to show the proportions, instead of the values. That's less appealing , but again, it could be done. The problem is which percentages do you show? The row percentages? The column percentages? There's no clear reason to prefer one over the other, and you cannot get the percentages without more than two pipes, and if you show the proportions, without reformatting, as in what I've done below, you still lose a lot of information with any of the three strategies. So you lost two points for doing this.

```
q14 %>% tabyl(solvent, diagnosis) %>%
  adorn_percentages(denominator = "row")
```

```
  solvent  definite   no  possible
 moderate 0.2500000 0.25 0.5000000
   modest 0.0000000 1.00 0.0000000
     none 0.5000000 0.00 0.5000000
 profound 0.6666667 0.00 0.3333333
```

```
q14 %>% tabyl(solvent, diagnosis) %>%
  adorn_percentages(denominator = "col")
```

```
  solvent definite  no possible
 moderate     0.25 0.5     0.50
   modest     0.00 0.5     0.00
     none     0.25 0.0     0.25
 profound     0.50 0.0     0.25
```

```
q14 %>% tabyl(solvent, diagnosis) %>%
  adorn_percentages(denominator = "all")
```

```
  solvent definite  no possible
 moderate      0.1 0.1      0.2
   modest      0.0 0.1      0.0
     none      0.1 0.0      0.1
 profound      0.2 0.0      0.1
```

You cannot (to my knowledge) show the actual percentages without adding a third pipe (to pull in the `adorn_pct_formatting` function), and that would invalidate your response. You also wouldn't be able to combine a table of proportions with rearranging the levels of solvent and diagnosis without more than two pipes, so that's not a good direction at all.

A call to the `cor` function or to `favstats` is incorrect, though, as neither `solvent` nor `diagnosis` is a quantitative variable. A call to the `chisq.test` function is also ineffective here, because it doesn't provide the table, just a test statistic and p value. Fitting a linear model and obtaining a summary is also a dead end in terms of failing to produce the table we need.

14.1.1 Grading 14

Question	Value	% Correct	% Points
14	4 points	60.0	62.1

- Running a table but using Q14 instead of `q14`, or `Solvent` instead of `solvent` cost you at least one point.
- If your code did not run for some other reason, like misspelling a variable name, or misplacing a character like a comma or parenthesis or something, you got no credit.
- Switching the order of the rows and columns was fine.
- Specifying marginal totals for only the rows or the columns was an odd choice, but we let it go.
- If you tried to pipe into, say, `xtabs` with the wrong pipe, you got no credit.
- Running a count without building a table, perhaps with the following code, got you 3/4 points.

```
q14 %>% group_by(solvent, diagnosis) %>% summarise(count=n())
```

```
`summarise()` regrouping output by 'solvent' (override with `.groups` argument)
```

```
# A tibble: 8 x 3
# Groups:   solvent [4]
  solvent  diagnosis count
  <chr>    <chr>     <int>
1 moderate definite      1
2 moderate no            1
3 moderate possible      2
4 modest   no            1
5 none     definite      1
6 none     possible      1
7 profound definite      2
8 profound possible      1
```

```
xtabs(~solvent + diagnosis, data = q14)
```

```
          diagnosis
solvent    definite no possible
  moderate        1  1        2
```

```
modest          0  1        0
none            1  0        1
profound        2  0        1
```

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.

15.1 Answer 15 is a line of code.

Again we could create a little table to be sure.

```
q15 <- tibble(exposure = c(2, 5, 42, 74, 77, 80, 89, 120, 130, 142),
              diagnosis = c("definite", "possible", "no", "no", "possible",
                            "possible", "definite", "possible", "definite",
                            "definite"))
```

I had in mind

```
mosaic::favstats(exposure ~ diagnosis, data = q15)
```

```
  diagnosis min    Q1 median  Q3 max  mean       sd n missing
1  definite   2 67.25  109.5 133 142 90.75 63.36863 4       0
2        no  42 50.00   58.0  66  74 58.00 22.62742 2       0
3  possible   5 59.00   78.5  90 120 70.50 47.86439 4       0
```

which seems the most direct way to get a numerical summary of the `exposure` distribution within each `diagnosis` type, which is what you needed.

Another way to produce the same result would be:

```
mosaic::favstats(~ exposure | diagnosis, data=q15)
```

and yet another way (which is a lot of work for no real gain) is:

```
q15 %>%
    group_by(diagnosis) %>%
     summarise(min = min(exposure), Q1 = quantile(exposure, 0.25),
               median = median(exposure), Q3 = quantile(exposure, 0.75),
               max = max(exposure), mean = mean(exposure),
               sd = sd(exposure), n = n(),
               missing = sum(is.na(exposure)))
```

If you wanted to use `summary`, you'd have to combine that with the use of the `by` function, in something like:

```
q15 %$% by(exposure, diagnosis, summary)
```

```
diagnosis: definite
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   2.00   67.25  109.50   90.75  133.00  142.00
------------------------------------------------------------
diagnosis: no
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     42      50      58      58      66      74
------------------------------------------------------------
diagnosis: possible
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    5.0    59.0    78.5    70.5    90.0   120.0
```

We were not looking for you to create a cutoff at 100 and apply it in this situation. That wouldn't tell us about the **distribution** of the `exposure` variable in sufficient detail.

Summarizing the data with a linear model predicting exposure using diagnosis would indirectly tell us the means of `exposure` in each group, but the mean alone isn't a sufficient summary for the distributions.

15.1.1 Grading 15

Question	Value	% Correct	% Points
15	4 points	52.9	53.6

- Lots of people tried to produce tables or tabyls, and that wasn't going to work.
- Others, instead of summarizing the variables we asked you to summarize, created other ones. That, too, wasn't going to work.
- We let you get away with naming the tibble `Q15` instead of `q15` but we probably shouldn't have, since capitalization changes are a common reason for R to throw an error.
- We didn't want you to include brackets for the code chunk, but didn't take points off if your answer was correct and you did.
- We didn't need you to use kable to specify rounding here, but we didn't add or subtract anything if you did so properly.
- Summarizing the mean alone wasn't enough, although it was about the only way to get partial credit on this question. You needed, at the least, to summarize a measure of center and a measure of spread.
- You can add `na.rm = TRUE` if you like, but it's unncessary.
- Inverting diagnosis and exposure's roles in `favstats` produces nothing of value, so that cost you all of the points.
- Neither of the two sets of code listed below work because they don't specify the tibble (`q15`):

```
mosaic::favstats(exposure ~ diagnosis)
```

and

```
group_by(exposure) %>%
    summarise(count = n(), mean(diagnosis), median(diagnosis)) %>%
    knitr::kable(digits = 2)
```

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`.

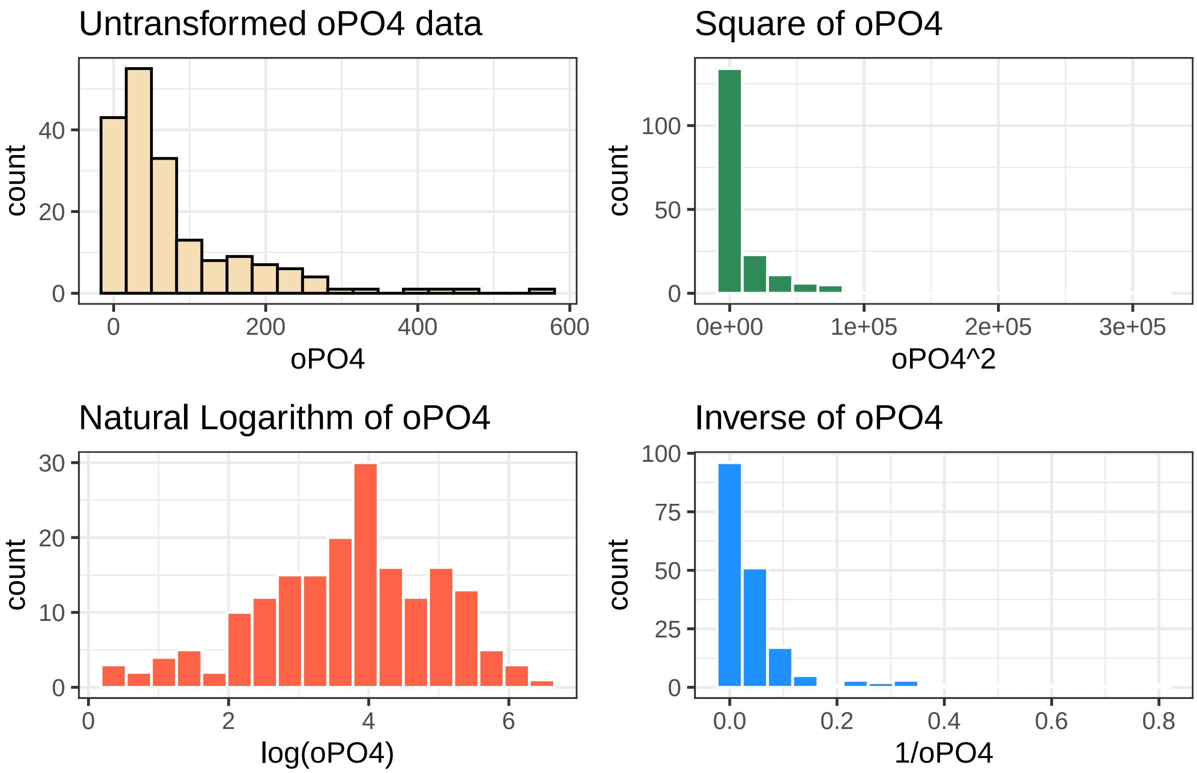


Figure for Question 16

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.

16.1 Answer 16 is b

The log transformation in this case yields a substantially closer fit to a Normal distribution than any of the other options, as we can see in the histograms, or if you like, in the Normal Q-Q plots below.

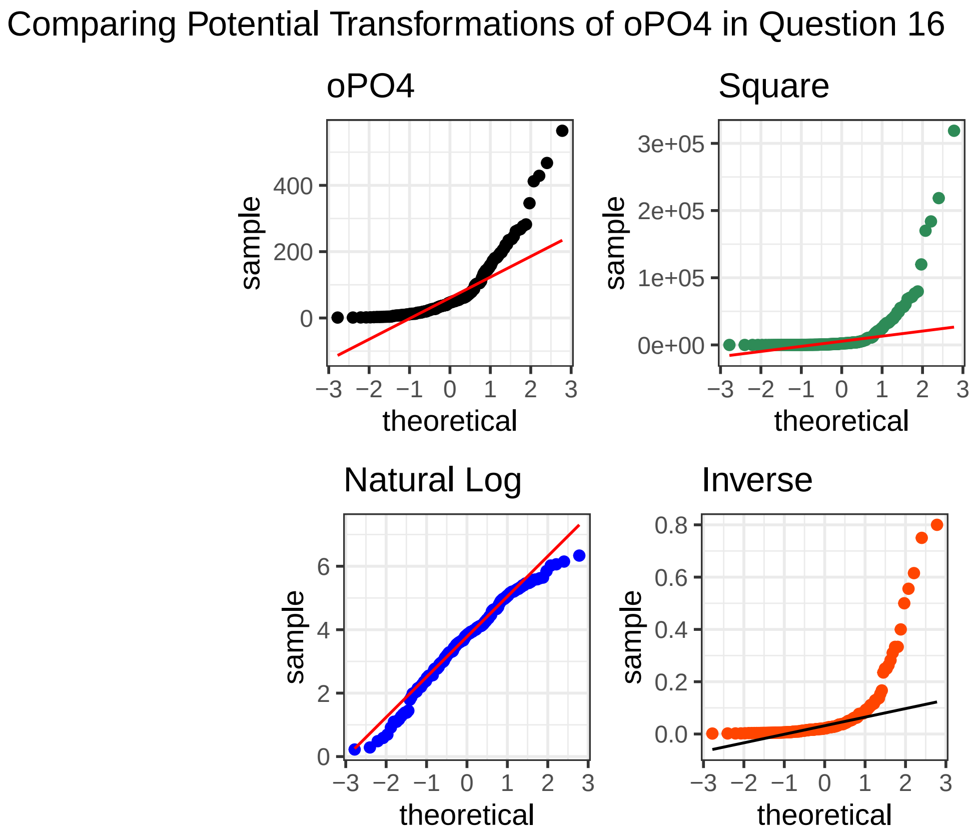
```r
p1 <- ggplot(algae, aes(sample = oPO4)) +
    geom_qq(col = "black") + geom_qq_line(col = "red") +
    theme(aspect.ratio = 1) +
    labs(title = "oPO4")

p2 <- ggplot(algae, aes(sample = oPO4^2)) +
    geom_qq(col = "seagreen") + geom_qq_line(col = "red") +
    theme(aspect.ratio = 1) +
    labs(title = "Square")

p3 <- ggplot(algae, aes(sample = log(oPO4))) +
    geom_qq(col = "blue") + geom_qq_line(col = "red") +
    theme(aspect.ratio = 1) +
    labs(title = "Natural Log")

p4 <- ggplot(algae, aes(sample = 1/(oPO4))) +
    geom_qq(col = "orangered") + geom_qq_line(col = "black") +
    theme(aspect.ratio = 1) +
    labs(title = "Inverse")

(p1 + p2) / (p3 + p4) + plot_annotation(title = "Comparing Potential Transformations of oPO4 in Questio
```



16.1.1 Grading 16

Question	Value	% Correct	% Points
16	3 points	≥ 95	≥ 95

- Most common incorrect response was e.

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?

17.1 Answer 17 is 162.

```
algae <- read_csv("data/algae.csv") %>%
    filter(complete.cases(a1, oPO4),
           a1 > 0)

nrow(algae)
```

```
[1] 162
```

17.1.1 Grading 17

Question	Value	% Correct	% Points
17	3 points	74.3	75.7

- I gave 1 point for 161 or 163.
- The most common incorrect response was 148.

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.

18.1 Answer 18 is 6.2

```
m1 <- lm(log(a1) ~ log(oPO4), data = algae)

augment(m1, newdata = tibble(oPO4 = 64))
```

```
# A tibble: 1 x 2
   oPO4 .fitted
  <dbl>   <dbl>
1    64    1.83
```

Remember that the prediction made by this model is of the natural logarithm of `a1` so we need to exponentiate this `.fitted` value to get our prediction for `a1`.

```
exp(1.828013)
```

```
[1] 6.221512
```

You might have preferred to calculate the result on your own. Let's look at the coefficients of the model.

```
tidy(m1)
```

```
# A tibble: 2 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)    4.31     0.195       22.1 2.26e-50
2 log(oPO4)     -0.597    0.0543     -11.0 2.84e-21
```

So the model is `log(a1) = 4.3098 - 0.5967 (log(oPO4))`. Now, if `oPO4` is 64, then `log(oPO4)` is 4.1588. So our predicted `log(a1)` would be…

`log(a1) = 4.3098 - 0.5967 (4.1588) = 1.8282`

and so our fitted `a1` value is `exp(1.8282)` or 6.2227, or, rounded to 1 decimal place, 6.2.

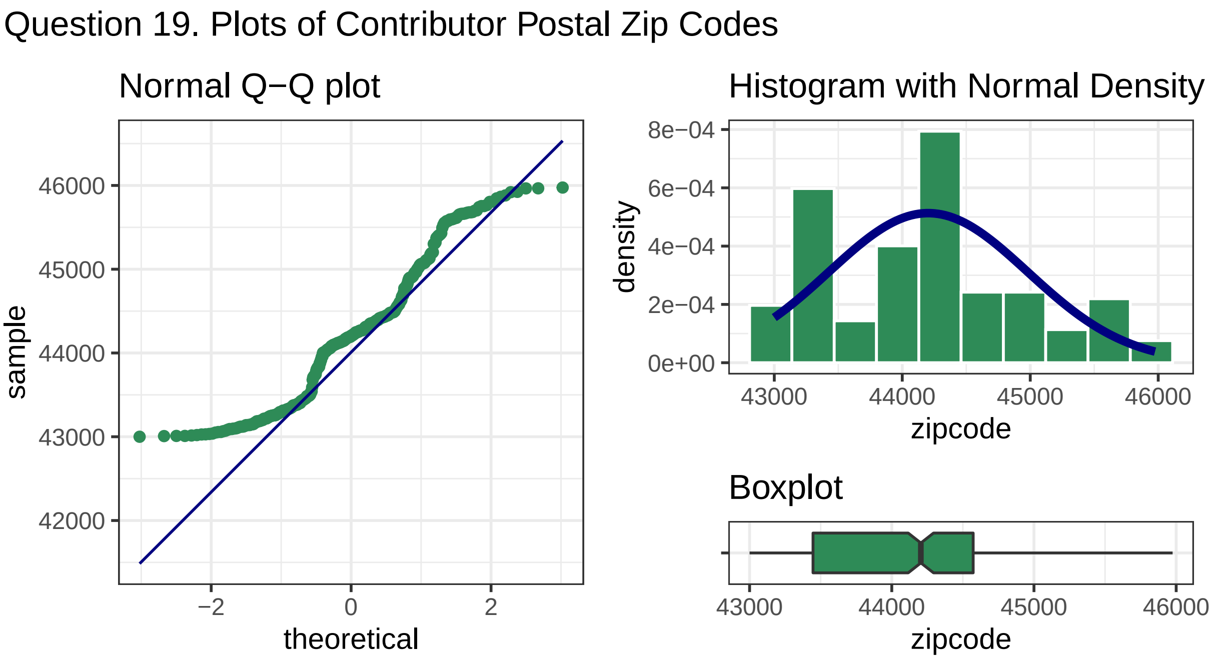
18.1.1 Grading 18

Question	Value	% Correct	% Points
18	3 points	41.4	41.4

This was one of the three questions I thought would be "trickiest" prior to you seeing the Quiz, and it definitely was difficult.

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.

19.1 Answer 19 is d.

Zip codes are numbers, but they aren't quantitative. Consider the CWRU Zip code of 44106. That has no units, and it's just a code. Instead, these are nominal categorical data. Of these four choices, only a **mode** could possibly be relevant. That would at least tell us which zip code had the most contributors in our recent sample.

19.1.1 Grading 19

Question	Value	% Correct	% Points
19	3 points	58.6	58.6

- Most common incorrect response was *b*, followed closely by **c**.
- Also selected: `e` and `a`.

This was one of the three questions I thought would be “trickiest” prior to you seeing the Quiz, but it went a little better than I’d thought it might.

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/m^2)	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.

20.1 Answer 20 is Round off the BMI values, a lot.

The table is already sorted alphabetically, and (suspiciously and conveniently) in decreasing order of body-mass index, so those elements are OK. Practice-level and day-specific averages are also provided for comparison. The names, incidentally, are all pitchers on the roster of the 2020 Cleveland Indians baseball team.

- Classifying people by their BMI into categories is actually a bad idea. You never want to lose granularity in a measured quantity like that deliberately. Sometimes it is forced on you, but creating categories is essentially never a good strategy, either when caring for people or caring for data.
- The table has a perfectly fine title already, and the headings are also understandable and reasonable.

20.1.1 Grading 20

Question	Value	% Correct	% Points
20	4 points	38.6	49.3

You were asked to suggest **a** worthwhile improvement in one sentence.

- People who wrote more than one sentence lost at least a point.
- People who suggested more than one improvement lost some credit, too, but retained some credit if one of their improvements involved rounding.

- If you suggested rounding the BMI down to 0, 1 or 2 decimal places or just to fewer decimal places, that was fine. If you thought 3 decimal places was still reasonable given the precision with which the height and weight were measured, you lost a point for that.
- If your response wasn't a complete sentence, you lost a point for that.
- Some people wanted to increase the volume of data in various ways. That wasn't what we were looking for.
- Some people wanted to add other statistical summaries, or color-coding, or de-identify the patients, or re-order the variables. None of these are meaningful improvements.
- Some people gave R code to try to round the data, but the data don't come from a tibble so the code suggested wouldn't work.

21 Question 21.

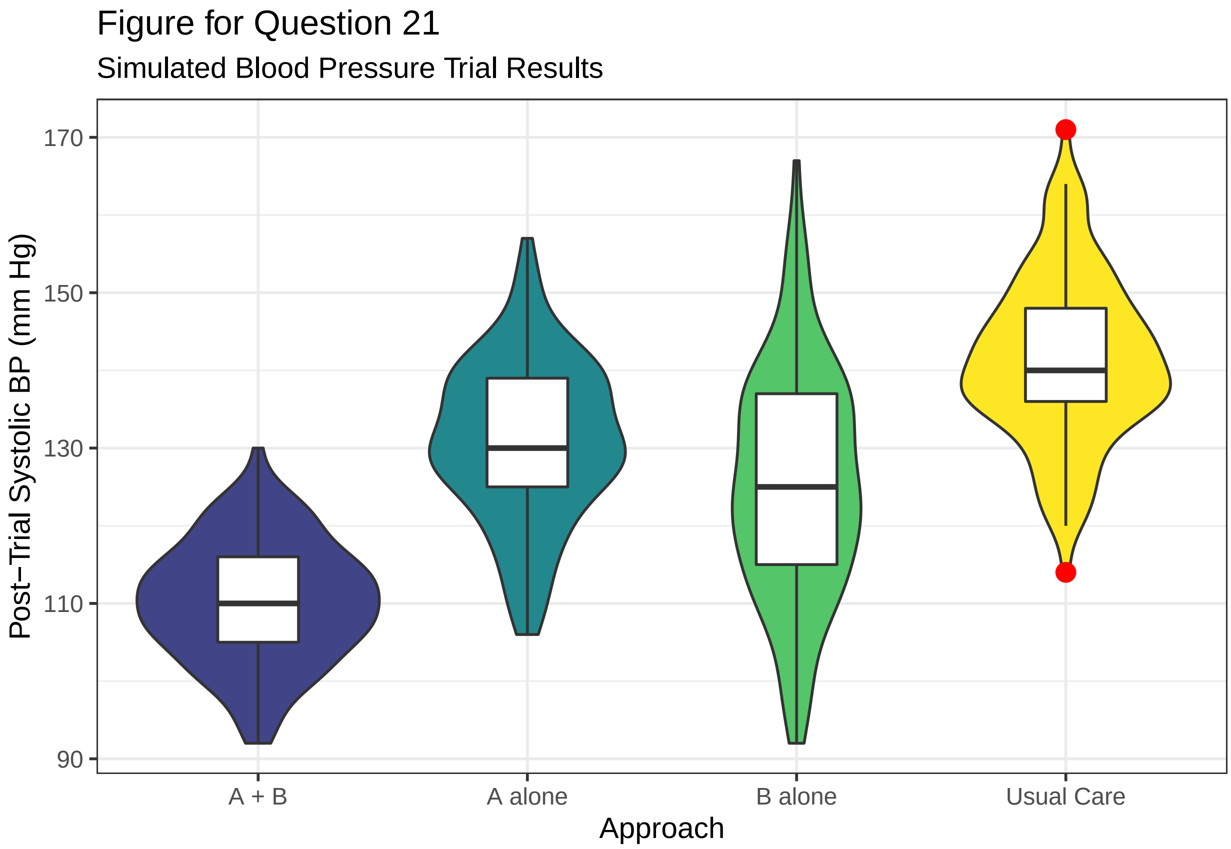
The initial results of the Systolic Blood Pressure Intervention Trial (SPRINT) received a lot of attention.

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.



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

21.1 Answer 21 is g.

All three statements are true, and this should be evident from the graph.

Statement I: Usual care has only one (outlier candidate) value below 120, while all other groups have their whiskers extending well below 120.

Statement II: The "B alone" group has clearly the largest spread, by its IQR or its whiskers.

Statement III: The 75th percentile for the A+B data is clearly below 120. 75% of 125 patients is 93.75 patients, so clearly more than 90 of the 125 A+B patients are below 120.

21.1.1 Grading 21

Question	Value	% Correct	% Points
21	3 points	78.6	92.9

In this question, you had 3 decisions to make (yes or no for each statement.) I gace one point for each correct decision.

The most common incorrect responses were `d` and then `f`.

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.

22.1 Answer 22 is e.

You can't tell, except that it might be either A+B or B. In fact, the A+B and B groups each have one patient with a post-trial SBP of 92 (which is the local minimum for each group), so there's a tie for the global minimum.

22.1.1 Grading 22

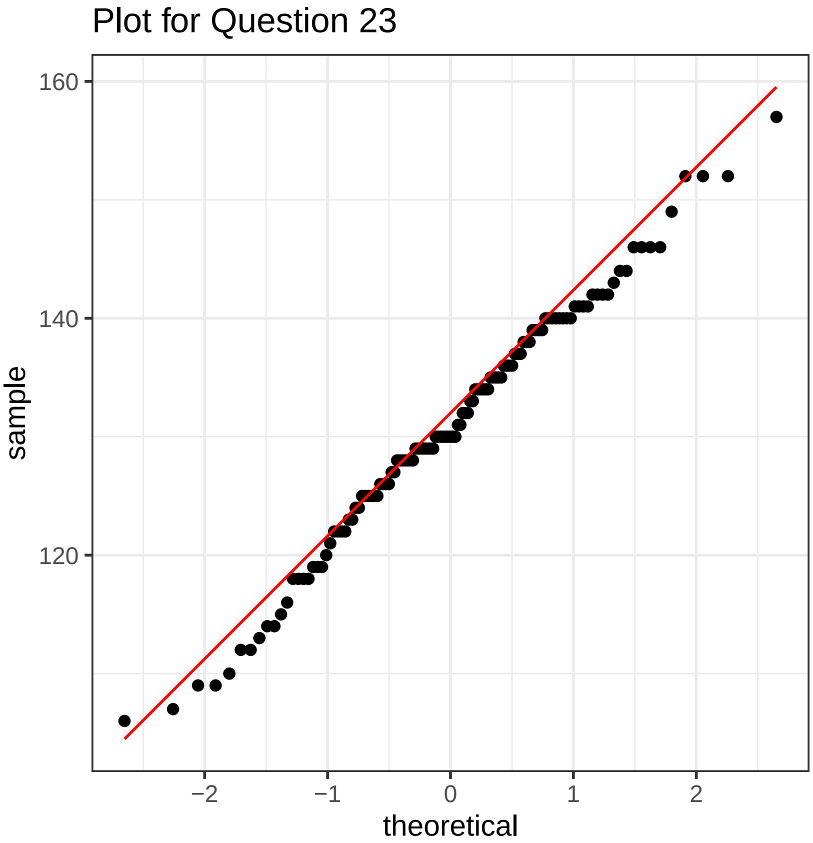
Question	Value	% Correct	% Points
22	3 points	78.6	78.6

- The most common incorrect responses were `d` and then `b`.
- People are often reluctant to choose "it's impossible to tell".

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

23.1 Answer 23 is a.

The normal Q-Q plot corresponds to the “A alone” group.

The easiest way to tell is to look at the range of the data, which, according to the Y axis of the Normal Q-Q plot should be from about 110 to 160. The only group meeting that standard is A alone.

23.1.1 Grading 23

Question	Value	% Correct	% Points
23	3 points	92.9	92.9

- The most common incorrect response was `d`.

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.)

24.1 Answer 24 is 4.

There are four characters who meet these requirements.

```
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
```

24.1.1 Grading 24

Question	Value	% Correct	% Points
24	4 points	94.3	94.3

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`?

25.1 Answer 25 is 18, but 20 is also a possible answer.

First, let's discuss the answer I intended, which was 18.

```
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
```

or perhaps you could first count the total number of characters...

```
starwars %>% nrow()
```

```
[1] 87
```

and then filter away the missing values to see what remains...

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

```
[1] 69
```

then subtract 87 - 69 to get 18.

On the day before the Quiz was due, a student brought up the interesting question of whether a value of "unknown" should be treated as missing in this Question. Let's explore that a bit.

- There are no characters in the `starwars` tibble whose `species` or whose `homeworld` is listed as "Unknown". (You can verify this for yourself, if you like.)
- There is 1 character whose `hair_color` is listed as "unknown".
- There are 3 characters whose `eye_color` is listed as "unknown", one of whom is also the character with unknown `hair_color`.

Who are those characters?

```
starwars %>%
  filter(hair_color == "unknown" | eye_color == "unknown") %>%
  select(name, species, homeworld, hair_color, eye_color)
```

```
# A tibble: 3 x 5
  name           species homeworld   hair_color eye_color
  <chr>          <chr>   <chr>       <chr>      <chr>
```

```
1 Ratts Tyerell  Aleena  Aleen Minor none       unknown
2 Wat Tambor     Skakoan Skako       none       unknown
3 Captain Phasma <NA>    <NA>        unknown    unknown
```

Now, Captain Phasma already has missing data for `species` and `homeworld`, but you could have argued that a character whose `eye_color` is unknown (Ratts Tyerell and Wat Tambor) was also "missing" one of these four pieces of information. So adding those two characters to the original list of 18 would give us a result of 20 characters.

I put in the Quiz 1 Note a statement that I considered "unknown" values in these variables to be non-missing values, making the answer 18. But you might have missed that note as I didn't post it until late in the process, so I would also have given full credit for the answer 20, if anyone had given it.

25.1.1 Grading 25

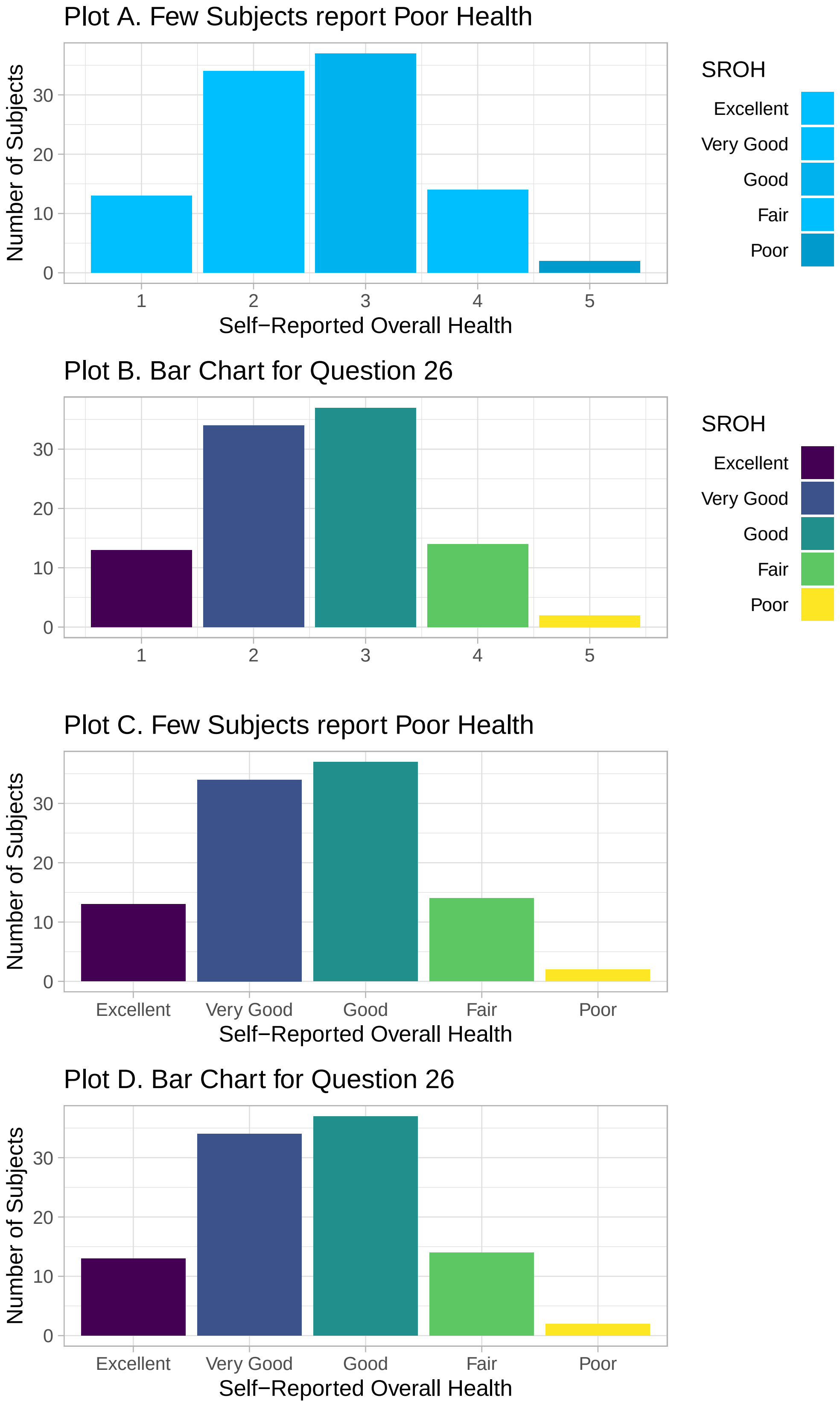
Question	Value	% Correct	% Points
25	4 points	77.1	77.1

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 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 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

26.1 Answer 26 is c.

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

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

26.1.1 Grading 26

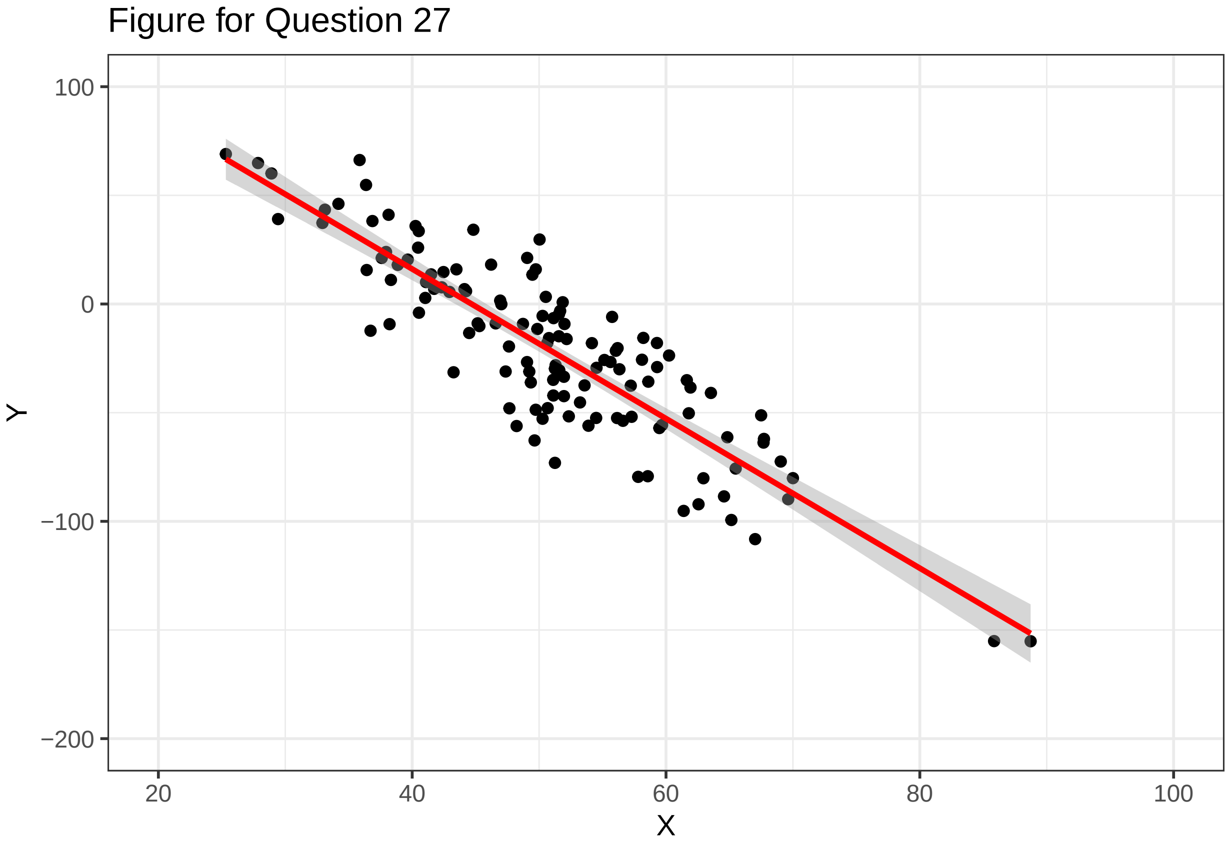
Question	Value	% Correct	% Points
26	3 points	91.4	91.4

- The most common incorrect responses were `d`, and then `a`.

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



27.1 Answer 27 is k.

- The models proposed in a, b, c, d, e and f all get the slope and intercept backwards
- R^2 cannot be negative so g and h are also incorrect.
- The Y-X slope is clearly negative (as X increases, Y decreases) so i and j are incorrect
- 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.

```
m1 <- lm(y ~ x, data = dat27)

tidy(m1) %>% select(term, estimate)
```

```
# A tibble: 2 x 2
  term        estimate
  <chr>          <dbl>
1 (Intercept)   154.
2 x              -3.44
```

```
glance(m1) %>% select(r.squared)
```

```
# A tibble: 1 x 1
  r.squared
      <dbl>
1     0.765
```

27.1.1 Grading 27

Question	Value	% Correct	% Points
27	3 points	88.6	88.6

- The most common incorrect responses were `g`. (R^2 cannot be negative.)
- Other selections were `c` and `l`.

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.

28.1 Answer 28 is all 9 of them.

See Chapter 9 of Leek's book.

28.1.1 Grading 28

Question	Value	% Correct	% Points
28	3 points	28.6	63.3

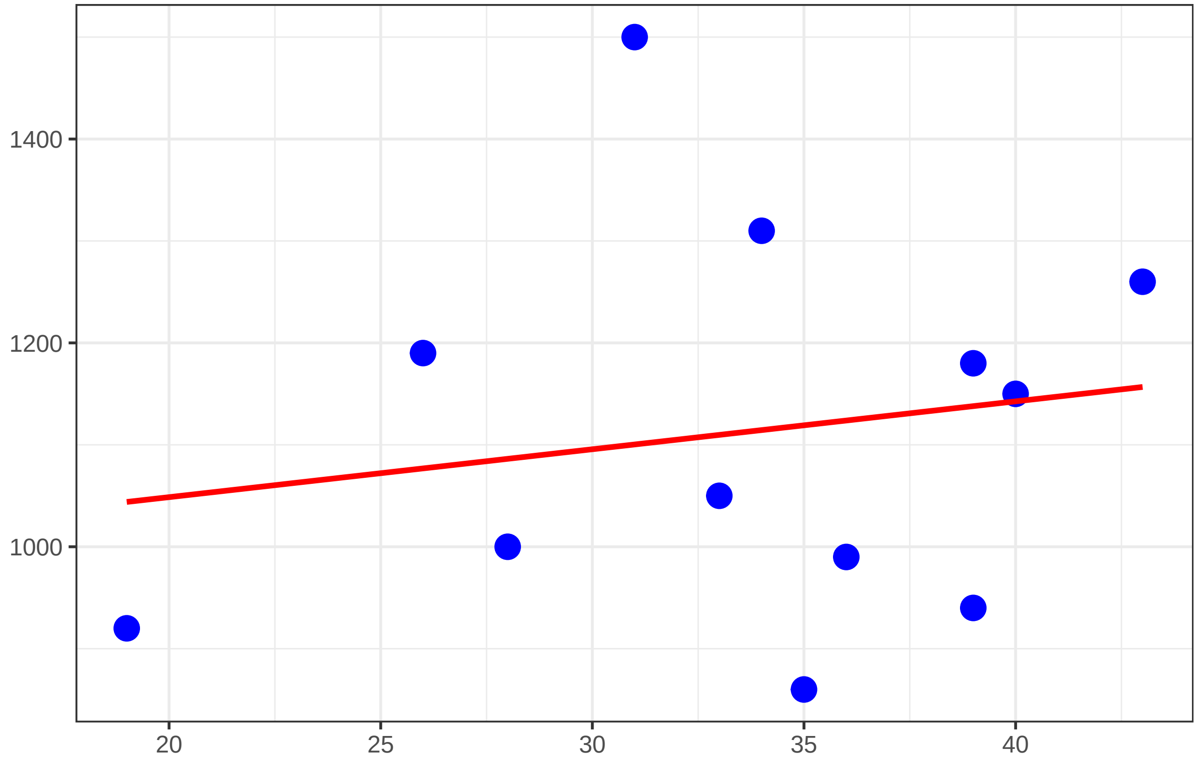
Here you had 9 little decisions (include or exclude) to make.

- I gave 2 points if you made 7 or 8 of them correctly.
- I gave 1 point if you made 6 of them correctly.
- The most common incorrect response left out `d`, but checked the rest.

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



29.1 Answer 29 is Add axis titles.

The correct response is to label the axes.

29.1.1 Grading 29

Question	Value	% Correct	% Points
29	4 points	61.4	87.9

You were asked to provide the **most important** change, in a sentence.

- The most common issue here was people who tried to provide multiple changes, one of which was to label the axes. Something like got you either 2 or 3 out of the 4 possible points, unless you clearly indicated that axis labeling was the **most** important thing and other things were not.
- Other things that people suggested weren't as important as labeling the axes. This included labeling the "outliers", modeling the data, say by adding a loess smooth, or adding a summary statistic, like a

correlation coefficient or the linear model's prediction equation3.

- Some people suggested adding scales to the x- and y- axes, or adding a linear model fit, but they're already there, so you lost a point or two for that.
- Similarly, the Figure has a title already, although it certainly could use some improvement.
- If you misspelled a word in your response (for example, label the "axis" instead of "axes" when referring to both of them) or wrote something that wasn't a sentence, then you lost a point or two.
- Some folks mentioned that the new axis labels should include the units of measurement. I agree, but didn't give or take away extra credit for this issue.

30 Overall Results

30.1 Table, by Item

Question	Value	% Correct	% Points
1	4 points	87.1	90.7
2	3 points	72.9	72.9
3 (I)	1 points	88.6	88.6
3 (II)	1 point	87.1	87.1
3 (III)	1 point	22.9	22.9
4	4 points	94.3	94.3
5	3 points	44.3	44.3
6	4 points	64.3	81.8
7	3 points	77.1	77.1
8	3 points	85.7	85.7
9	3 points	≥ 95	≥ 95
10	3 points	81.4	81.4
11	3 points	70.0	70.0
12	3 points	48.6	48.6
13	4 points	≥ 95	≥ 95
14	4 points	60.0	62.1
15	4 points	52.9	53.6
16	3 points	≥ 95	≥ 95
17	3 points	74.3	75.7
18	3 points	41.4	41.4
19	3 points	58.6	58.6
20	4 points	38.6	49.3
21	3 points	78.6	92.9
22	3 points	78.6	78.6
23	3 points	92.9	92.9
24	4 points	94.3	94.3
25	4 points	77.1	77.1
26	3 points	91.4	91.4
27	3 points	88.6	88.6
28	3 points	28.6	63.3
29	4 points	61.4	87.9

31 Calculating Your Score

- Add up your points on the 29 items. That's your **raw score**.
- Then add five points. That's your **score on Quiz 1**.

31.1 Results and Dr. Love's interpretation

Quiz 1 Score	# of students	How Dr. Love thinks of that score
95 to 102	12	Outstanding, a high A
90 to 94	10	Excellent, a solid A
85 to 89	10	Very Good, a low A
80 to 84	11	Good, a high B
75 to 79	9	Fine, a solid B
70 to 74	6	OK, a low B
below 70	12	Needs to improve

Each of you will receive an email from me with details on your score.

32 Please, don't panic!

There is no reason for any of you to panic in light of your score on the Quiz. If you didn't do as well as you'd hoped, take a look at the answer sketch and see what you missed. Then concentrate on the work you're doing now and going forward, not on your score on the Quiz. Improvement matters.