Fall 2021 Midterm I

The exam is open book. This means you can use electronic or hard copies of all class materials and can use datahub or a local version of R/Rstudio if you wish. You may not use the internet to search for the answers or to inform your answers. Using the internet is strictly prohibited and any evidence of this may result in a 0 on the exam.

While you take the exam, you are prohibited from discussing the test with anyone. If you are taking the test after your classmates, you are also prohibited from talking to them about the test before you take it. Evidence of cheating may result in a 0 on the exam and be reported to the Student Conduct Board.

Berkeley's code of conduct is here: https://sa.berkeley.edu/code-of-conduct. See Section V and Appendix II for information about how UC Berkeley defines academic misconduct. In particular note the sections on cheating and plagiarism.

UC Berkeley Honor Code
"As a member of the UC Berkeley community, I act with honesty, integrity, and respect for others." Please carefully read the statements below, and indicate your understanding and intent to adhere to the UC Berkeley Honor code by typing your name in the space below. I agree not to engage in any of the following behaviors:

- Copying or attempting to copy from others during an exam or on an assignment.
- Communicating answers with another person during an exam.
- Pre-programming a calculator or other personal electronic device to contain answers, or using other unauthorized information for exams.
- Using unauthorized materials, i.e. prepared answers.
- Allowing others to do an assignment or a portion of an assignment for you, including the use of a commercial term-paper service.
- Submitting the same assignment for more than one course without prior approval of all the instructors involved.
- Collaborating on an exam or assignment with any other person without prior approval from an instructor.
- Taking an exam for another person or having someone take an exam for you.
- Altering a previously graded exam or assignment for the purpose of a grade appeal or of gaining points in a re-grading process.
- Submitting an electronic file the student knows to be unreadable or corrupted instead of a completed assignment.

Type your name and SID below.

Name:

Enter your name:

Enter your SID:

INSTRUCTIONS:

1.Use Adobe Reader or Acrobat as a stand-alone application (NOT in a browser) to complete this assignment. This software can be accessed for free for UCB students **here**

2.Give your responses ONLY in the space provided. Do NOT add any additional text boxes.

3.Please rename the file LASTNAME_FIRSTNAME_Midterm1_Spring2022.pdf

4.The exam is due on February 17th by 12pm noon SHARP. We recommend submitting by 11:45am at the latest to ensure your submission is received. **Be sure to upload your submission to the correct VERSION of the midterm I takehome portion on Gradescope.** Submitting your exam to the incorrect version on Gradescope will result in an automatic 50% reduction in points. We will also not accept any late exams under any circumstances.

NOTES:

- Unless otherwise specified in the question, format your answers according to the following guidelines:
  - present your answers rounded to two decimal places
  - present proportions as % values (40.50% rather than .405)
- All logs are natural log base e

MAKE SURE YOU ARE WORKING WITH THIS DOCUMENT IN ADOBE AND YOU ARE NOT IN A BROWSER WINDOW

Problem 1: 4 points

Problem 2: 9 points

Problem 3: 1 point

Problem 4: 6 points

Problem 5: 2 points

Problem 6: 6 points

Problem 7: 8 points

Total: 36 points

Question 1 [4 points total]

Suppose you have a dataset, `wildfire_US`, that has 6 variables: `fire_name`, `state`, `county`, `acres_burned`, `num_hospitalized`, and `pop_size`. The states in the `state` variable are capitalized and written out in their entirety (e.g., Alabama rather than alabama or AL) and the `num_hospitalized` and `pop_size` variables represent thousands of people (e.g., 100 means 100,000 people).

1.a. [1 point] Write a *single* line of code (which can contain many functions) to output a subset of data that contains only California state and 4 of the 6 original columns (state, number of acres burned, number of people hospitalized, and population size) and assign it to the object `wildfire_subset`.

SOLUTION: wildfire_subset <- wildfire_US %>% select(state, acres_burned, num_hospitalized, pop_size) %>% filter(state == "California")

1.b. [2 points] You want to visualize the distribution of the actual number of people hospitalized in your `wildfire_subset` dataset. Use 2 or 3 sentences (not lines of code) to describe how you would do this.

SOLUTION: Create a new column that is the original variable, `num_hospitalized`, multiplied by 1000 to get the exact number of people hospitalized. Create a histogram of the new column of the exact number of people hospitalized.

1.c. [1 point] You notice that there are 2 peaks in the plot you created, so you decide to create separate plots of the number of people hospitalized based on whether the `acres_burned` were greater than or equal to 1000 acres. After creating a new binary variable called `ge_1000`, what line of code would you use to create these separate plots?

a) `ggplot(data = wildfire_subset, aes(x = num_hospitalized)) + geom_histogram() + facet_grid( ~ ge_1000)`
b) `ggplot(data = wildfire_subset, aes(x = num_hospitalized)) + geom_histogram() + facet_wrap( ~ ge_1000)`
c) `ggplot(data = wildfire_subset, aes(x = num_hospitalized)) + geom_bar() + facet_grid( ~ ge_1000)`
d) `ggplot(data = wildfire_subset, aes(x = num_hospitalized)) + geom_bar() + facet_wrap( ~ ge_1000)`

Answer: [ ]

SOLUION: b) `ggplot(data = wildfire_subset, aes(x = num_hospitalized)) + geom_histogram() + facet_wrap( ~ ge_1000)`

Question 2 [9 points total]

Background: A hospital readmission is an episode when a patient who has been discharged from a hospital is admitted again within a specified time interval. We define this time interval between the first discharge and readmission as *time-to-readmission*. For example, if patient *A* was first discharged from the hospital on September 6th and then admitted again on September 15th, then we say that the *time-to-readmission* of patient *A* is 9 days.

2.a. [1 point] Suppose you are a researcher in the local public health department and would like to study hospital readmission among diabetes patients. You have collected previous patient information including age, sex, income level, type of diabetes, glucose level, and time-to-readmission. You want to create a model using this previously collected patient information to determine future patients' time-to-readmission. What type of research question are you trying to address in this study?

- ◯ Descriptive
- ◯ Etiologic/Causative
- ◯ Predictive
- ◯ None of the above

SOLUTION: c) Predictive

2.b. [2 points] Below are the names of the variables in your dataset and their descriptions:

Variable name	Description
`age`	age (numeric)
`sex`	biological sex (factor: female, male)
`income`	income level (factor: low, middle, high)
`d_type`	type of diabetes (factor: 1, 2)
`glucose`	glucose level (numeric)
`time_to_readmission`	time interval between first discharge and readmission (in days)

You would like to determine whether there is a difference in the average time-to-readmission for patients with type 1 or type 2 diabetes. Fill in the blanks below to find the average readmission times by diabetes type.

```
patient_data <- patient_data %>% __blank1__(__blank2__) %>% __blank3__(mean_readmit = __blank4)
```

A: ☐
B: ☐
C: ☐
D: ☐

SOLUTION: - blank 1: group_by - blank 2: d_type - blank 3: summarize - blank 4: mean(time_to_readmission)

2.c. [2 points] You notice that individuals with type 2 diabetes tend to have a lower time-to-readmission on average, meaning they are readmitted more quickly than those with type 1 diabetes. You create an experiment to try to improve the average readmission time of type 2 diabetes patients by randomizing half of the type 2 diabetes individuals from your original sample into a new diabetes management program while the other half get standard care. What is the factor variable being manipulated in this study and how many levels are there for this factor?

- ○ type of diabetes
- ○ number of patients
- ○ average readmission time
- ○ enrollment in diabetes management

How many levels are there for this factor?

SOLUTION: d) enrollment in diabetes management; 2 levels

2.d. [2 points] How many treatment groups are there? Are there always the same number of treatment groups as factor levels? Explain in one sentence.

SOLUTION: 2 treatment groups: diabetes management or no diabetes management/ No, there are not always the same number of treatment groups as factor levels. For example, in lecture we talked about photoperiods and light wavelengths and flowering. There were 4 levels of photoperiods and 3 levels of light wavelengths, but there were 12 treatments (4*3 = 12 combinations)

2.e. [1 point] Upon completion of your study, you notice that the individuals randomized to the experimental treatment had improved readmission times. Is it likely that another third variable was influencing this relationship? Why or why not?

SOLUTION: No, it is not likely. Randomization decreases likelihood of confounding

2.f. [1 point] We believe our sample of individuals in the experiment are representative of the general population of individuals with type 2 diabetes. Thus, our study is _____ valid.

SOLUTION: externally

Question 3 [1 point total]

For each of the given scenarios, select which `geom_` function would be most appropriate to visualize the given variables in the dataset.

3.a. [0.5 point] An intern at the San Francisco Public Health Department receives a dataset called `covid_tracker` that contains the number of individuals diagnosed with COVID-19 and their age group (child, adolescent, adult). You want to make a plot using both of these variables.

a) `geom_point()`
b) `geom_histogram()`
c) `geom_bar()`
d) `geom_abline()`

Answer: [ ]

SOLUTION: c) geom_bar()

3.b. [0.5 point] Chandler wants to see if there is a trend between the number of Piazza posts and the number of issues with Datahub for students in PH 142. You want to make a plot with both of these variables.

a) `geom_point()`
b) `geom_histogram()`
c) `geom_bar()`
d) `geom_abline()`

Answer: [ ]

SOLUTION: geom_point()

Question 4 [6 points total]

Gigi P. Lot is a professor at the UC Berkeley School of Public Health with a fascination of the `ggplot` package in R. For her research project, she is working with a dataset that contains information on patients' demographic and heart vitals information (https://www.kaggle.com/johnsmith88/heart-disease-dataset). The data dictionary for the dataset is provided below:

Attribute	Description
age	Age of patient in years
sex	Sex of patient (0 = female; 1 = male)
trestbps	Resting blood pressure (in mm Hg on admission to the hospital)
chol	Serum cholestoral in mg/dl
fbs	Fasting blood sugar > 120 mg/dl (0 = False; 1 = True
thalach	Maximum heart rate achieved
exang	Exercise induced angina (1 = yes; 0 = no)
slope	Slope of the peak exercise ST segment
ca	Number of major vessels (0-3) colored by flouroscopy
target	Presence of heart disease in the patient (0 = No disease; 1 = Disease)

```
heart <- read_csv("chandler_mt1_qs/heart.csv")
```

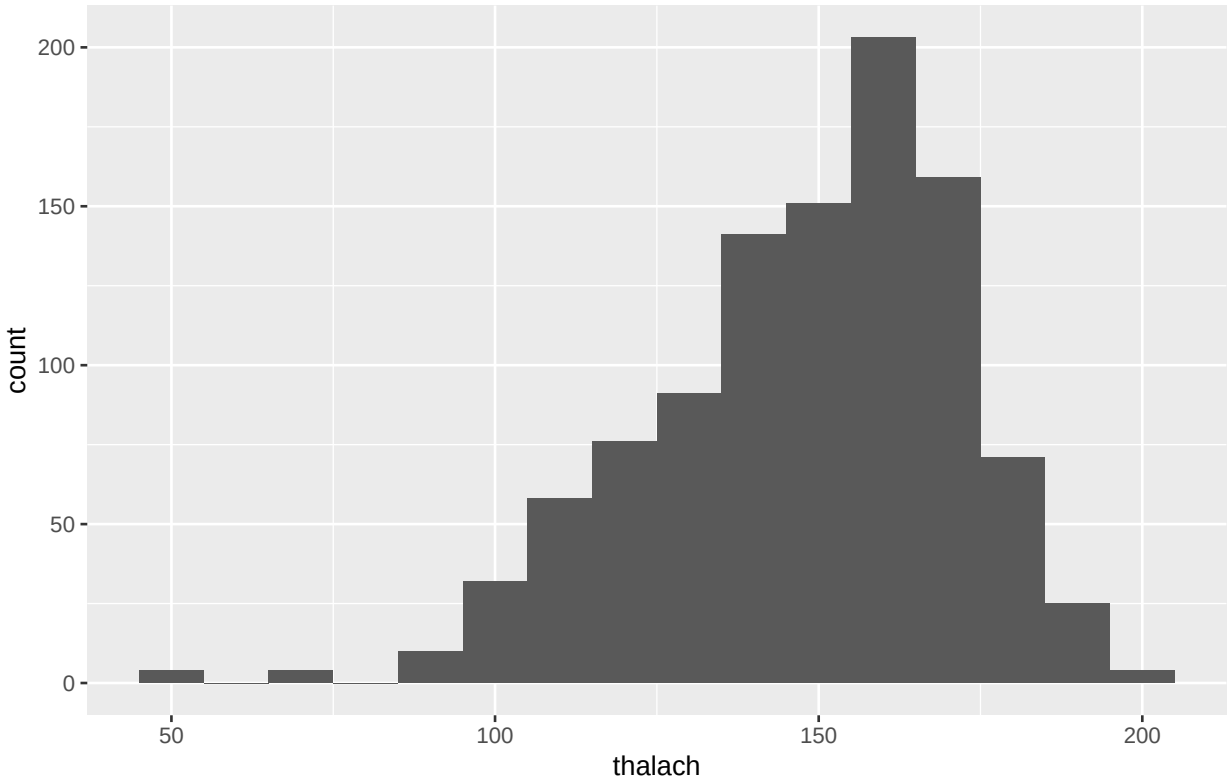
```
## Rows: 1029 Columns: 14
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (14): age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpea...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
heart <- heart %>% rbind
heart <- heart %>% select(-cp, -oldpeak, -restecg, -thal)
heart$sex <- as.character(heart$sex)
```

4.a. [1 point] You want to join Professor P. Lot in her research, but she would like to test your ability to visualize information using `ggplot` before letting you join her. Which of the following lines of code best recreates the graph below?

```
q6i <- ggplot(heart, aes(thalach)) +
  geom_histogram(binwidth = 10)
ggsave("q6i_hist.png")
```

```
## Saving 9 x 4.5 in image
```


200
150
100
50
0
count
50
100
150
200
thalach

a) `ggplot(heart, aes(thalach)) + geom_histogram(binwidth = 20)`
b) `ggplot(heart, aes(thalach)) + geom_histogram(binwidth = 10)`
c) `ggplot(heart, aes(thalach)) + geom_bar()`
d) `ggplot(heart, aes(thalach)) + geom_bar(stat = "identity")`
e) `ggplot(heart, aes(thalach)) + geom_point()`

Answer:

SOLUTION: b) `ggplot(heart, aes(thalach)) + geom_histogram(binwidth = 10)`

4.b. [2 points] Describe the distribution of the data (plotted above) in terms of four characteristics of distributions.

SOLUTION: - shape: skewed left

- center: somewhere between 140-150, unimodal
- spread: range is from around 48 to 205
- outlier: potential outlier at around 45

4.c. [1 point] How does the mean of the distribution above compare to the median?

- ◯ mean > median
- ◯ median > mean
- ◯ mean = median
- ◯ cannot determine from the plot

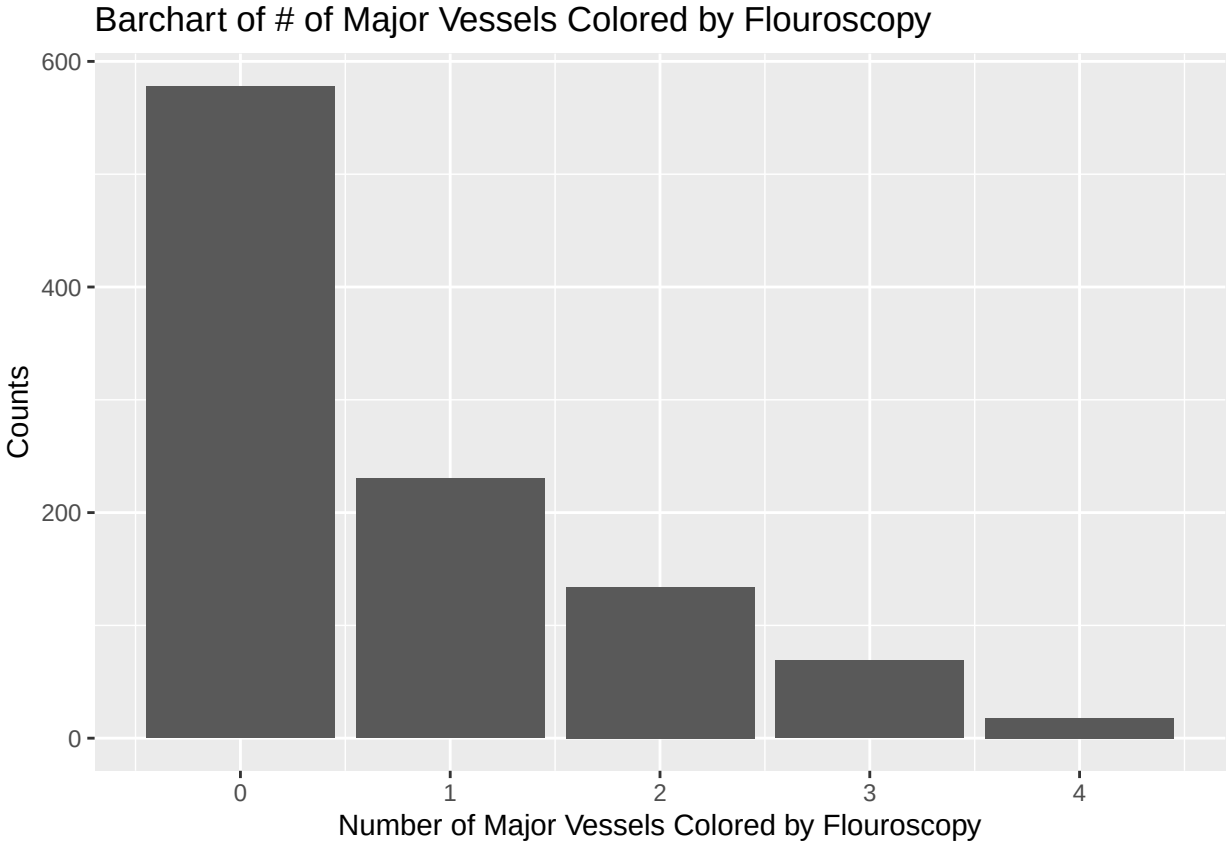
SOLUTION: b) median > mean

4.d. [2 points] The professor welcomes you to her team after completing the above tasks. She aggregates the dataset and gives you the following dataframe, called `heart_agg`:

```
## # A tibble: 5 x 2
##   major_vessels count
##           <dbl> <int>
## 1             0   578
## 2             1   230
## 3             2   134
## 4             3    69
## 5             4    18
```

She then asks you to visualize the new aggregated data. Provide the code for the visualization below, making sure to add the following labels:

- title: "Barchart of # of Major Vessels Colored by Flouroscopy"
- x-axis: "Number of Major Vessels Colored by Flouroscopy"
- y-axis: "Counts"



SOLUTION: ggplot(heart_agg, aes(major_vessels, count)) + geom_bar(stat = 'identity') + labs(title = "Barchart of # of Major Vessels Colored by Flourouscopy", x = "Number of Major Vessels Colored by Flouroscopy", y = "Counts")

Question 5 [2 points total]

Suppose you have a dataset called `lifeExp_data` with data on 1000 individuals' life expectancy in years, `lifeExp`, as well as their `BMI`, `age`, and `sex`. You want to see if BMI is an accurate linear predictor of life expectancy.

5.a. [1 point] Complete the line of code below to fit the linear regression model between BMI and life expectancy.

```
lm(__________, data = lifeExp_data)
```

SOLUTION: lifeExp ~ BMI

5.b. [1 point] You plot the data and the line of best fit and notice a point that deviates from from the rest of the data in *both* the x and y directions. Which of the following statements is true?

◯ The data must be nonlinear since we have an outlier, so linear regression is not an appropriate way to model BMI and life expectancy.

◯ The r^2 value is likely very small (close to 0) since we have an outlier.

◯ Removing the point from the dataset would likely affect the resulting regression line.

◯ The point is an influential point, not an outlier.

◯ All of the above

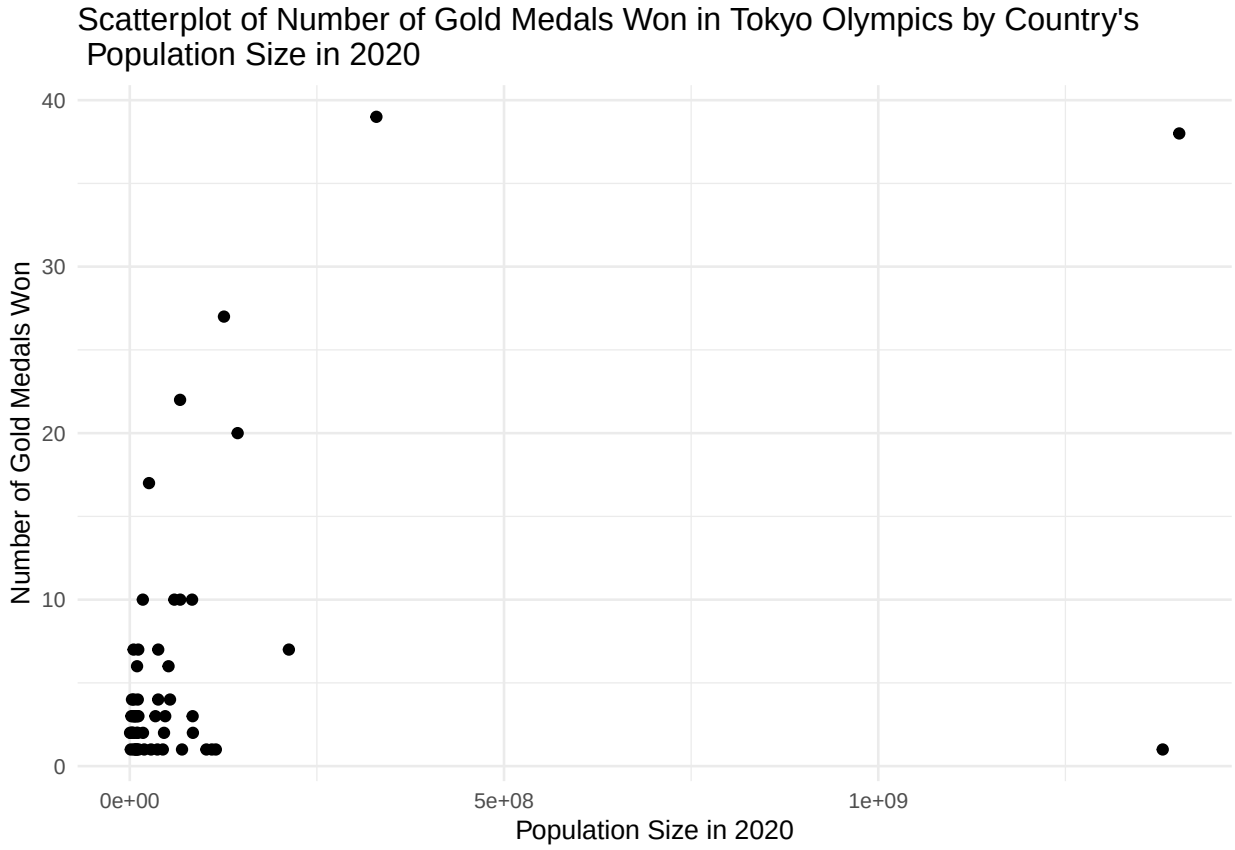
SOLUTION: c) Removing the point from the dataset would likely affect the resulting regression line.

Question 6 [6 points total]

A friend approaches you and begins to brag that their home country won more gold medals in the most recent Tokyo Olympics than your home country. You politely remind them that the population of your home country is much smaller than theirs, but your friend insists that population has no correlation with the number of gold medals won. They point out that New Zealand and Brazil won the same number of gold medals and Brazil has a population 40 times greater than New Zealand. You walk away fuming, but then remember that you can try to use regression to show that your friend is incorrect.

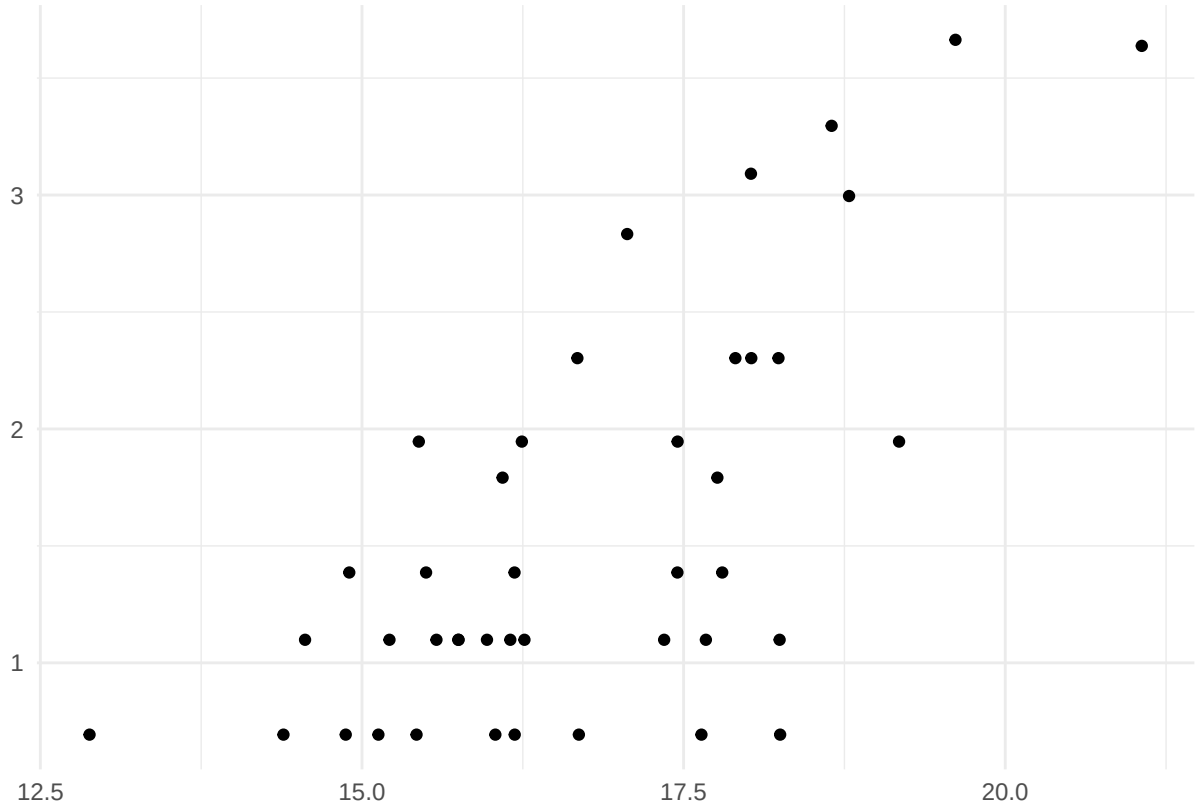
6.a. [1 point] The plot below is the result of plotting `gold_count` against `pop_2020`. Is linear regression an appropriate way to model the relationship between the number of gold medals and the country's population size in 2020? Why or why not?

```
## Rows: 77 Columns: 9
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): team, NOCCode
## dbl (7): rank, pop_2020, gold_count, silver_count, bronze_count, total_count...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```



SOLUTION: No, linear regression is not appropriate. Many of the points are clumped in the bottom left and there are outliers that would pull the line of best fit upwards (away from the other data points at the bottom), meaning a line would not be the best representation of the data.

6.b. [1 point] You transform your data and plot them in the scatterplot as seen below. What transformation(s) did you perform, and on which variable(s)? How do you know?



SOLUTION: Natural log transformations on both x and y variables (the number of gold medals and the population size) Can describe: - both axes being different in the transformed plot compared to original - picking a point on the original plot and seeing if their predicted - transformations match the transformed plot - example: (7 medals, 2.5e8 population size). log(7) = 1.95 and log(2.5e8) = 19.3. On the transformed plot you can find a point at around (1.95, 19.3) so you know you logged both the number of gold medals and the population size.

6.c. [1 point] Describe the relationship of the transformed data as seen in the scatterplot in 6.b. You do not have to comment on outliers, but do comment on the other characteristics of linear relationships we discussed in lecture.

SOLUTION: - strength: weak or moderate - form: linear - direction: positive

6.d. [1 point] You run a regression on the data from 6.b. and are given the output below. Interpret the slope in the context of this problem.

```
## # A tibble: 2 x 5
##   term                estimate std.error statistic     p.value
##   <chr>                  <dbl>     <dbl>     <dbl>       <dbl>
## 1 (Intercept)            -4.56     1.05      -4.33 0.0000986
## 2 pop_2020_transformed    0.366    0.0626     5.84  0.000000785
```

SOLUTION: For every one unit increase in the natural log of the population, the average number of the natural log of gold medals won increases by 0.366.

6.e. [1 point] Suppose that an unknown country has a population size of 521,000,000. Which of the following is the country's predicted number of gold medals won in the 2021 Olympics?

- ◯ 20.1 gold medals
- ◯ 1.67 gold medals
- ◯ 2.78 gold medals
- ◯ 16.2 gold medals

SOLUTION: d) 16.2 log(gold_count) = -4.559 + 0.366*log(pop_2020) log(gold_count) = -4.559 + 0.366*log(521000000) log(gold_count) = 2.78 gold_count = 16.2

6.f. [1 point] You look at the r^2 value to see how well the regression model fits the data. What is the correct interpretation of the r^2 value?

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic     p.value    df logLik   AIC   BIC
##       <dbl>         <dbl> <dbl>     <dbl>       <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1     0.460         0.447 0.643      34.1 0.000000785     1  -40.0  86.1  91.3
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

a) About 46.0% of our data will fall close to our regression line.
b) About 46.0% of the variance of the transformed `gold_count` variable can be explained by the transformed `pop_2020` variable.
c) We need approximately 46.0% more data in order to make an accurate regression line.
d) We cannot interpret r^2; we must interpret r: 67.8% of our data will fall close to our regression line.
e) We cannot interpret r^2; we must interpret r: 67.8% the the variance of the transformed `gold_count` variable can be explained by the transformed `pop_2020` variable.

Answer:

SOLUTION: b) About 46.0% of the variance of the transformed `gold_count` variable can be explained by the transformed `pop_2020` variable.

Question 7 [8 points total]

Researchers were interested in looking at the relationship between COVID-19 vaccination rates and age group. The following data looks at a random sample of 1850 individuals, considering their vaccination status and age group (0-30 years old, 31-60 years old, or 61+).

7.a. [2.5 points] Fill in the blanks in the following two-way table.

	Vaccinated	Unvaccinated	Total
0-30	590	110	A
31-60	B	100	640
61+	360	C	510
Total	D	E	1850

A: [ ]
B: [ ]
C: [ ]
D: [ ]
E: [ ]

SOLUTION: A = 700, B = 540, C = 150, D = 1490, E= 360

7.b. [1 point] What is the marginal distribution of vaccination in this sample? Round your answer to 2 decimal places.

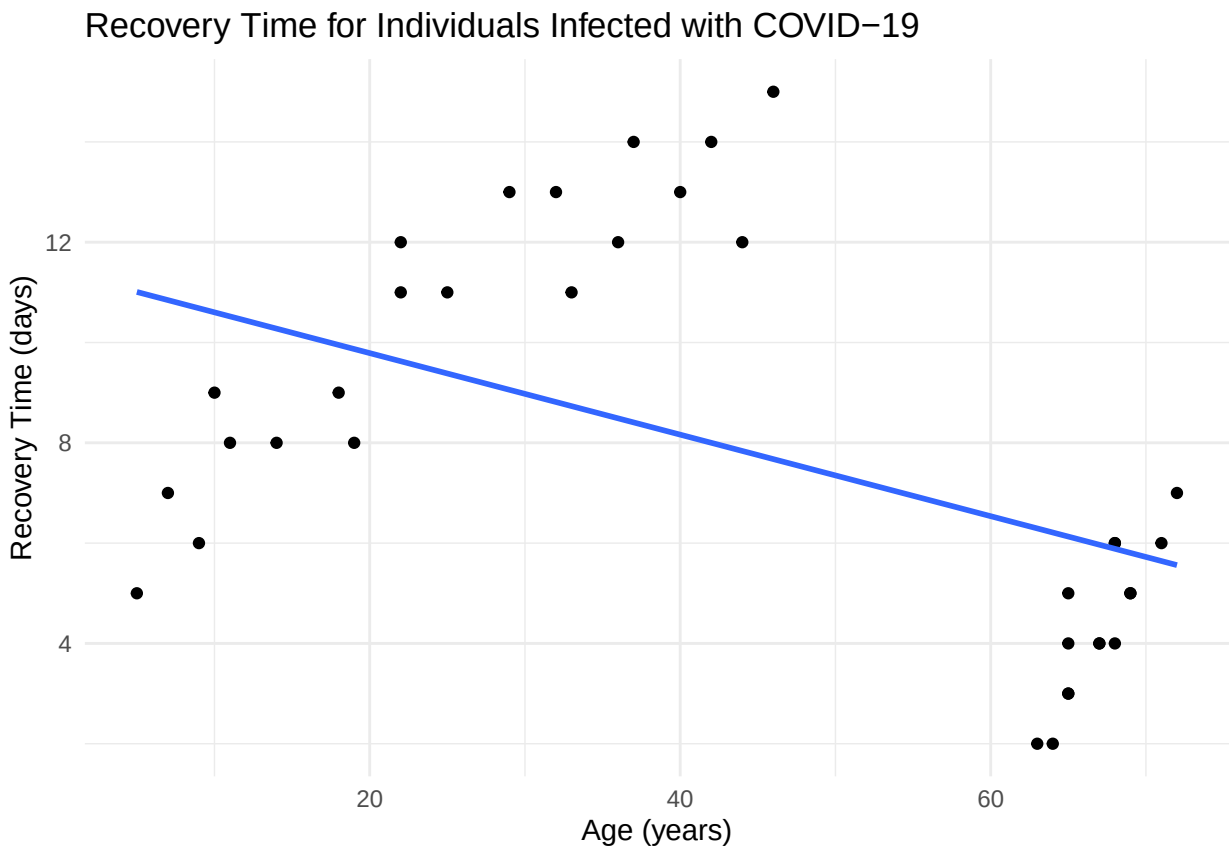
SOLUTION: The percentage of vaccinated people is 80.54% (1490/1850). The percentage of unvaccinated people is 19.46% (360/1850).

7.c. [1 point] What is the conditional distribution of vaccination among those who are 61+? Round your answer to 2 decimal places.

SOLUTION: The vaccinated percentage among 61+ people is 70.59% (360/510). The unvaccinated percentage among 61+ people is 29.41% (150/510).

7.d. [0.5 point] The researchers also collected data on the recovery time for individuals who were recently infected with COVID-19. The plot below shows the recovery time plotted against age a few weeks after the COVID-19 vaccine was released to older individuals in January 2021. How would you describe the overall relationship between X and Y if you only considered the regression line (the line of best fit)? Make sure your answer mentions the actual X and Y variables.

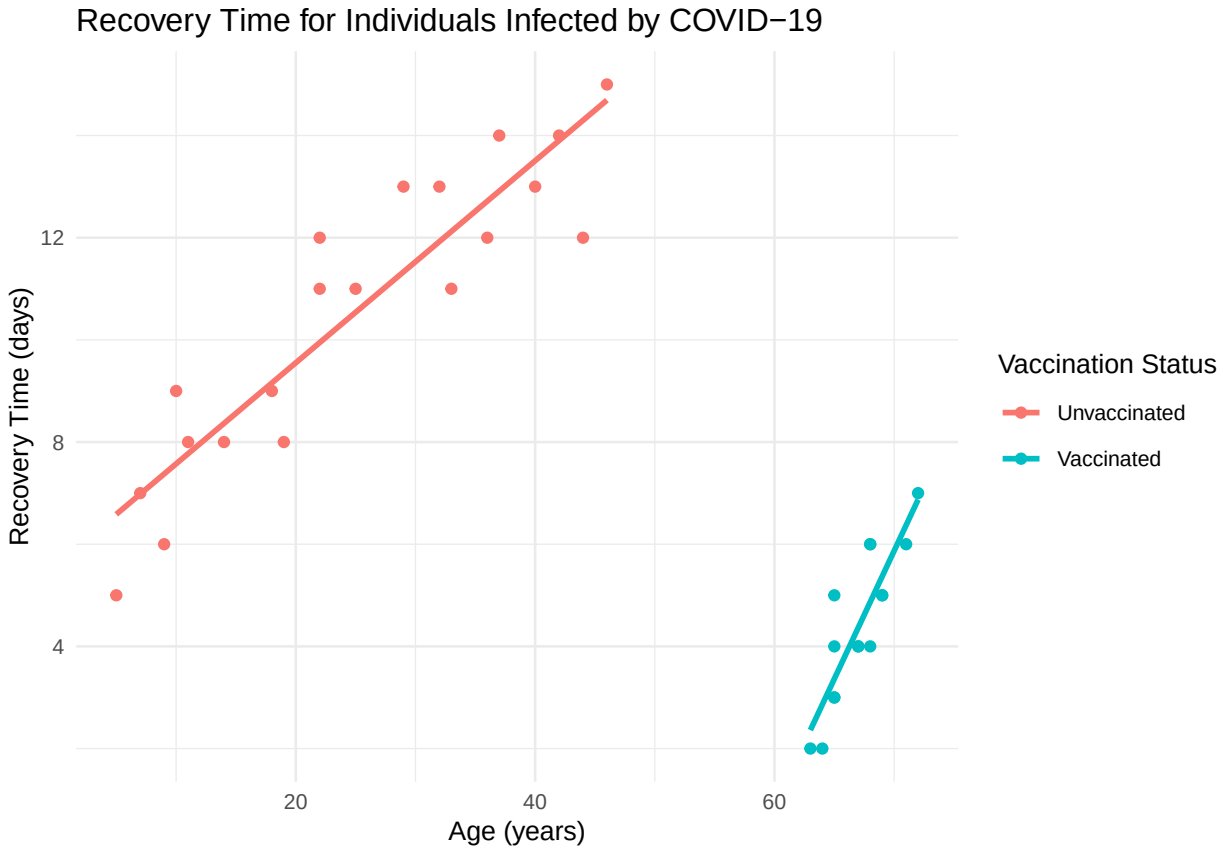
```
## `geom_smooth()` using formula 'y ~ x'
```



SOLUTION: As age increases, the recovery time decreases.

7.e. [1 point] The researchers notice that there are 2 distinct groups of data in the plot above. They group the individuals by vaccination status and re-plot the data below. Now, how would you describe the relationship between X and Y now that the data is stratified by vaccination status? (Again, make sure your answer mentions the variables under study.)

```
## `geom_smooth()` using formula 'y ~ x'
```



SOLUTION: Recovery time increases with increasing age for both vaccinated and unvaccinated individuals.

7.f. [2 points] The researchers are baffled by their findings. In one or two sentences, name this phenomenon and describe why this might occur.

SOLUTION: Simpson's Paradox. The vaccine has been made available to people 65+. So older folks (who have longer recovery times on average) are now also vaccinated, and vaccine is reducing their recovery time. This explains why the overall recovery time decreases while within the stratified vaccination groups, the recovery time still increases with age.