Data C8 Fall 2024 — Foundations of Data Science — Final

Berkeley Honor Code [1 point]

"As a member of the UC Berkeley community, I act with honesty, integrity, and respect for others."

Initials: ________________

PRELIMINARIES

Full Name: ______________________________

Student ID Number: ______________________________

Name of person to your left: ______________________________

Name of person to your right: ______________________________

Exam Location (Building & Room Number): ______________________________

TA's name (write *N/A* if in self-service lab): ______________________________

Date & Time: 8:10-11:00am, Monday, December 16, 2024

INSTRUCTIONS

- Do **not** open the exam until you are instructed to do so.
- Write your **initials at the top of each page**.
- There are **7 questions** and **19 pages** on this exam, including this cover page. **Read the instructions and point values carefully** for each question, part and subpart.
- Parts within a question may or may not depend on each other! Keep this in mind if you are stuck on a question part. You should see if you can answer the next part.
- Where relevant, you may assume that all necessary Python modules have been imported.
- You may only have with you: a pencil(s), an eraser(s), your student ID, a water bottle, and your final reference sheet, unless you have received pre-approved accommodations otherwise.
- If you need to use the restroom, bring your phone, exam, reference sheet, and student ID to the front.

MULTIPLE CHOICE QUESTION TYPES

For questions with **circular bubbles**, you should fill in exactly *one* choice. **Please fill in completely.**

○ You must choose either this option

○ Or this one, but not both!

For questions with **square boxes**, you may fill in *multiple* choices. **Please fill in completely.**

□ You could select this choice.

□ You could select this one too!

1 Multiple Choice [38 points]

Read the directions carefully for each part and subpart.

a. (7 points) Prof Jeremy is working with a table which contains both categorical and numerical variables.

(i) (4 points) The following are examples of procedures or tasks that we have discussed throughout the course. Which are part of the Exploration aspect of data science? *Select all that apply.*

√ **Calculating correlation coefficients between pairs of numerical variables**

□ Calculating a confidence interval for a population parameter

□ Performing a hypothesis test

□ Creating a graph of averages using two numerical variables

√ **Identifying the units of variables in the dataset**

√ **Visualizing a categorical variable**

□ Predicting the value of a variable belonging to a new observation

√ **Determining whether the table should be considered a sample or a population**

(ii) (3 points) Prof Jeremy wants to investigate a potential linear relationship between two particular numerical variables in the table, x and y. In which of the following situations below would he conclude that there is not a linear relationship between x and y? *Select all that apply.*

□ He calculates the correlation coefficient between the two variables to be -0.90.

√ **He visualizes the variables with a scatter plot and sees that as x increases, y does not change.**

√ **He fits a regression model to explain y with x, examines the corresponding residual plot and sees a pattern in the residuals.**

b. (3 points) Dagny attempts to predict a student's final score on a test (out of 100) using the number of hours they have studied and their midterm score (out of 100) as predictor variables. What techniques below can she use? *Select all that apply.*

□ k-Nearest neighbors classification

√ **k-Nearest neighbors regression**

√ **Multiple linear regression**

□ Bayes' classifier

c. (5 points) Consider using a sample to construct a 95 percent bootstrap confidence interval for a population parameter. In which situations would we have less than 95 percent confidence that the interval captures the parameter? *Select all that apply.*

√ **When the sample is not representative of the population.**

√ **When the sample size is small.**

√ **When the parameter of interest is the population maximum.**

d. (5 points) True or False: When evaluating the performance of a prediction model, one should determine whether predictions are accurate for the data points that were used to develop the model.

○ True

√ **False**

e. (4 points) For which of the following tables discussed in the text/lecture did we pretend that we did not know the value of a population parameter we were interested in? *Select all that apply.*

√ **SF Compensation**

□ Mothers and newborns

√ **United Airlines**

□ Top grossing movies

□ Breast cancer diagnosis

□ Alameda County jury panels

f. (4 points) Sam performs an A/B test on a dataset and rejects the null hypothesis that groups A and B come from the same underlying distribution. What must be true for her to make the claim that being in group A versus group B causes the difference?

○ Her p-value cutoff must be equal to or lower than 5%.

√ **Her data must be collected in a mechanism/context that allows her to make a causal claim from it.**

○ She has made a visualization of the distributions between the two groups that shows a clear difference.

○ She needs a large enough sample size.

g. (4 points) Cai has two categorical distributions in a table and is testing the null hypothesis that they come from the same population versus the alternative hypothesis that they come from different populations. She is using the total variation distance as her test statistic, and creates an array called `simulated` that contains the simulated test statistics. She stores the observed value of the test statistic into `observed`.

(i) (3 points) What hypothesis are the test statistics simulated under?

√ **Null** ○ Alternative ○ Either is possible.

(ii) (1 point) Which of the following Python expressions evaluates to the p-value? *Select all that apply.*

□ `np.count_nonzero(simulated >= observed)`
√ `np.average(simulated >= observed)`
□ `np.count_nonzero(simulated <= observed)`
□ `np.average(simulated <= observed)`

h. (3 points) How many rows and columns will `your_table` have once the following Python code runs?

```
your_table = Table().with_columns(‘colors’, make_array(‘blue’,‘red’,‘light blue’,‘purple’),
                                  ‘favorite?’, make_array(False, True, False, False))

your_table.where(‘colors’, are.containing(‘blue’)).drop(0)
```

○ 4 rows and 1 column ○ 2 rows and 1 column ○ 3 rows and 2 columns

○ 3 rows and 1 column √ **4 rows and 2 columns** ○ 2 rows and 2 columns

i. (3 points) Oski is attempting to predict the price of a book sold at the Cal Student Store using the weight of the book.

The mean price of the books the Store has data on is $y = 100$ dollars and the mean weight of the books is $x = 5$ pounds. Oski uses linear regression to estimate y with x and then makes two 90% percent prediction intervals: one for the price of a book having weight $x = 6$ pounds and the other for the price of a book weighing $x = 2$ pounds. Which interval will be wider?

√ **The interval for a book weighing $x = 2$ pounds** ○ The interval for a book weighing $x = 6$ pounds

2 Welcome Survey [17 points]

At the beginning of the Fall 2024 semester, Battle Bus University, which has a student population of 50,000, required all students to fill out a welcome survey that gave the admissions office a glimpse into the background of the student body. The Data 8 staff has been working with Battle Bus University and obtain a random sample of 1,000 of these students. The staff stored them in a table called `welcome` that includes two of the questions that were asked to the students.

- **Extroversion** *(integer)*: How extroverted does the student feel, on a scale of 1 to 10? Higher scores indicate greater levels of extroversion.
- **Sleep** *(integer)*: How many hours does the student sleep on a typical night?

Extroversion	**Sleep**
4	8
7	8
6	7
7	7

... (996 rows omitted)

a. (5 points) Brandon would like to make a confidence interval for the population median extroversion (for all Battle Bus University students) using the `welcome` table. He sets up the following function to obtain 100 bootstrap medians; part of the function is detailed below. Fill in the blank corresponding to each line of code.

```
# 1   def bootstrapped_medians():
# 2      stats = ___(a)___
# 3      for i in np.arange(___(b)___):
# 4           new_sample = welcome.sample(k = ___(c)___, with_replacement = ___(d)___)
# 5           new_median = percentile(___(e)___, new_sample.column('Extroversion'))
# 6           ___________________(f)_________________________
         return stats
```

(i) Fill in blank (a):

Solution: `make_array()`

(ii) Fill in blank (b):

Solution: `100`

(iii) Fill in blank (c):

Solution: `1000`

(iv) Fill in blank (d):

Solution: `True`

(v) Fill in blank (e):

Solution: 50

(vi) Fill in blank (f):

Solution: `stats = np.append(stats, new_median)`

b. (4 points) Brandon has run the completed version of `bootstrapped_medians()` and saved his results into `welcome_medians`. Fill in the skeleton below to compute an 80 percent confidence interval for the population median extroversion score.

```
# 1            lower = percentile(___(g)___, welcome_medians)
# 2            upper = percentile(___(h)___, welcome_medians)
# 3            interval = make_array(lower, upper)
```

(i) Fill in blank (g):

Solution: 10

(ii) Fill in blank (h):

Solution: 90

c. (3 points) The confidence interval was computed to be [5, 7]. Which of the following are true statements relating to the confidence interval you just constructed? *Select all that apply.*

- √ **If Brandon's friend Marissa repeats this process 500 times, she can expect that roughly 400 of the confidence intervals she calculates will contain the true population median.**
- □ 80% of all Data 8 students have extroversion scores between 5 and 7.
- √ **The original sample median extroversion score Brandon found in the data could have been 4.**
- □ There is an 80 percent probability that the population median extroversion score is between 5 and 7.
- √ **A 99% confidence interval calculated using the sample will be wider than this 80% confidence interval.**

d. (3 points) The Squirrels Of Battle Bus University want to analyze the `welcome` data as well. They take two random samples: one of size 750, and the other of size 1000, and create an 80 percent confidence interval for the population average hours of sleep using each one. How does having only 750 students versus 1000 students in the sample affect the width of an interval estimate of a population parameter?

- ○ The width of the interval calculated with 750 students will be smaller than the width of the interval calculated with 1000 students.
- √ **The width of the interval calculated with 750 students will be larger than the width of the interval calculated with 1000 students.**
- ○ The sample size does not affect the width of an interval estimate.

e. (2 points) Arfa uses the Squirrels' sample of size 1000 to calculate a 99% confidence interval for the population average amount of hours slept, using the Central Limit Theorem. Which of the following are true statements that could help Arfa compute the interval? *Select all that apply.*

- □ Regardless of the population distribution of hours slept, the distribution representing the sample amounts of hours slept will be roughly normal.
- √ **Regardless of the population distribution of hours slept, the distribution representing the distribution of sample mean amounts of hours slept will be roughly normal.**
- √ **The distribution of sample mean amounts of hours slept will be balanced at the population mean amount of hours slept.**
- □ According to Chebyshev's bounds, approximately 99% of the hours slept will lie within three standard deviations of the mean amount of hours slept. We will use these bounds to create the interval.

3 Election Day in Oz [19 points]

The citizens of the fictional land of Oz are interested in which candidate will win the position of Wizard in an election that takes place every four years. The table `Oz` contains the following columns:

- **Year** *(integer)*: The election year. Years included: 1980, 1984, 1988, ..., 2020.
- **Satisfaction** *(integer)*: The percentage of respondents in the poll nearest to the election date who answered *Satisfied* to the question: *In general, are you satisfied or dissatisfied with the way things are going in Oz at this time?*
- **Votes** *(integer)*: The number of votes won by the Current Wizard.

Year	**Satisfaction**	**Votes**
2020	28	232
2016	37	227
2012	33	332
2008	13	173

...(7 rows omitted)

The table below shows the value of several statistics calculated using the full `Oz` table above. You may use these values to help answer the questions that follow:

Statistic	Value
Correlation between Satisfaction and Votes	0.68
Mean of Satisfaction	36
Median of Satisfaction	37
SD of Satisfaction	15
Mean of Votes	278
Median of Votes	266
SD of Votes	127

a. (3 points) Write out the **slope** of the least squares regression line for Votes, using Satisfaction as a predictor. **Do not simplify your arithmetic**.

Solution: The slope of the least squares regression line is given by:

$$\frac{0.68 * 127}{15}$$

b. (1 point) Write out the **intercept** of the least squares regression line for Votes, using Satisfaction as a predictor. **Do not simplify your arithmetic**.

Solution: The intercept of the best least squares regression line is given by:

$$\left(278 - \frac{0.68 * 127}{15} * 36\right)$$

c. (3 points) True or False: You could have used numerical optimization to obtain a slope and intercept similar to what you obtained using the table of statistics given.

√ **True** ○ False

d. (2 points) Once simplified, the least squares regression line is roughly:

$$\text{estimate of Votes} = 6 * \text{Satisfaction} + 64$$

Interpret the slope of the line in the context of the problem.

○ For every six percentage points increase in satisfaction amongst the surveyed citizens of Oz, the current Wizard is expected to receive one more electoral vote.

○ For every percentage point increase in satisfaction for an individual citizen of Oz, the current Wizard is expected to receive six more electoral votes.

√ **For every percentage point increase in satisfaction amongst the surveyed citizens of Oz, the current Wizard is expected to receive six more electoral votes.**

○ For every six percentage points increase in satisfaction for an individual citizen of Oz, the current Wizard is expected to receive one more electoral vote.

e. (2 points) Below is data on the 2024 election in a one-row table. This table is separate from the `Oz` table.

Year	**Satisfaction**	**Votes**
2024	26	226

In the box below, state in **one sentence** whether the least squares regression line overestimates or underestimates the 2024 results, and by how many electoral votes. Show your work.

Solution: 6*(26) + 64 = 220

The residual for 2024: 226 - 220 = 6 votes.

The line underestimated the number of votes by 6.

f. (2 points) Write an arithmetic expression which evaluates to the Euclidean distance between the new 2024 row and the 2020 row in the `Oz` table. Consider only the variables **Satisfaction** and **Votes**, and **do not simplify**.

Solution: The expression is given by

$$\sqrt{(28-26)^2 + (232-226)^2}$$

g. (2 points) The Current Wizard needs at least 270 votes to win the election.

(i) (1 point) Define a function called `election_winner` below that takes in a number, `votes`, as its argument, returns 'Current Wizard' if at least 270 votes have been won and returns 'Opponent' if less than 270 votes have been won.

Solution:

```
def election_winner(votes):
   if votes >= 270:
        return "Current Wizard"
   else:
        return "Opponent"
```

(ii) (1 point) Fill in the blanks (a) and (b) to add a column called **Winner** to the `Oz` table which, for each election cycle, reads 'Current Wizard' if at least 270 votes were won by the Current Wizard and 'Opponent' if the Current Wizard won less than 270 votes.

```
Oz = Oz.____(a)____(____(b)____)
```

(i) Fill in blank (a):

Solution: `with_columns`

(ii) Fill in blank (b):

Solution: `'Winner', Oz.apply(election_winner, 'Votes')`

h. (4 points) Below is a full and updated version of the Oz table which includes a column called **Distance** containing the Euclidean distance between each row and the 2024 row, rounded to the nearest integer, as well as the incomplete **Winner** column added onto the table.

Year	Satisfaction	Votes	Distance	Winner
2020	28	232	6	Opponent
2016	37	227	11	
2012	33	332	106	
2008	13	173	54	
2004	44	286	62	
2000	62	266	53	
1996	39	379	153	
1992	22	168	58	
1988	56	426	202	
1984	48	525	299	
1980	19	49	177	

(i) (2 points) Using the 3-nearest neighbors **classification model**, predict the winning party for 2024.

○ Current Wizard √ **Opponent**

(ii) (2 points) Using the 3-nearest neighbors **regression model**, predict the winning party for 2024.

○ Current Wizard √ **Opponent**

4 This Question is About Burritos [20 points]

After complaining about how expensive it is to eat out near the Berkeley campus, Data 8 staff members are interested in understanding if there is a linear relationship between the distance from the UC Berkeley campus (in miles) and the price of restaurant items. Edwin, an avid burrito consumer, takes a sample 500 of restaurants in East Bay serving carne asada burritos and stores it in the `carne` table below.

- **Location** *(string)*: Name of the restaurant, with the specific location supplied after a hyphen in the case that the restaurant is a chain.
- **Price** *(float)*: The price of a carne asada burrito, in U.S. dollars.
- **Distance** *(float)*: The distance from Evans Hall on the Berkeley campus, in miles.

Location	Price	Distance
Chipotle - Shattuck	11.35	0.64
La Burrita - Southside	11.29	0.44
Tacos Sinaloa	13	0.39
Las Cabañas	13.99	0.91
Tacos Mi Rancho	15	5.94
La Casa de Maria	10.5	17.73

...(494 rows omitted)

a. (5 points) Edwin would like to get a sense of the center and spread of the distribution of prices of burritos for restaurants in the `carne` dataset. Which of the following techniques he might use? *Select all that apply.*

□ classification

□ regression

✓ **data visualization: a histogram**

□ data visualization: a scatter diagram

✓ **creating a table of summary statistics**

□ confidence interval

□ hypothesis test

b. (3 points) Tiffany wants to use the `slope(table, x, y)` function we have defined in class to compute the slope of the regression line explaining price using distance, but she accidentally passes in a data table where the `x` and `y` columns are in standard units instead of original units. Will the function still accurately compute the slope in standard units?

✓ **Yes** ○ No

c. (4 points) After looking at the slope of Tiffany's regression line, Thomas claims that the true relationship between the distance from campus and the price of burritos in the East Bay is governed by a downward sloping line, and that the data in our table are generated by taking points from this line and pushing them up or down at random. However, Thomas's friend Jack argues that the true line is flat, and there is no association between price and distance.

Before conducting any inferential procedures, which of the following must the staff do? *Select all that apply.*

□ The staff should split the data into training and testing sets.

□ The staff should standardize their data.

✓ **The staff should verify that the sample has been collected at random.**

d. (3 points) Thomas and Jack decided they wanted to perform a hypothesis test to figure out if there is a negative linear relationship between distance and price, and they have come up with a few options for their null and alternative hypotheses.

1. There is a negative linear association between the distance from campus and the price of carne asada burritos; the slope of the true line is less than 0.
2. There is no linear association between the distance from campus and the price of carne asada burritos; the slope of the true line is 0.
3. There is no linear association between the distance from campus and the price of carne asada burritos, the intercept of the true line is 0.
4. There is a negative linear association between the distance from campus and the price of carne asada burritos, the intercept of the true line is less than 0.

(i) (2 points) Which of the option(s) above are valid null hypotheses? *Select all that apply.*

□ 1 √ **2** □ 3 □ 4

(ii) (1 point) Which of the option(s) above are valid alternative hypotheses? *Select all that apply.*

√ **1** □ 2 □ 3 □ 4

e. (5 points) Ella computes a 95% confidence interval for the slope of the true line as part of the hypothesis test and gets

$$[-0.05, -0.015]\,.$$

.

(i) (2 points) Based on the interval computed above, which of the possible p-value cutoffs would lead Ella to reject the null hypothesis? *Select all that apply.*

□ 0.01 √ **0.05** √ **0.10**

(ii) (3 points) Assume that, based on the interval above and her chosen p-value, that Ella rejected the null hypothesis. Which of the following statements can she make? *Select all that apply.*

√ **The evidence suggests a negative linear relationship between the distance from the UC Berkeley campus and the price of a carne asada burrito in East Bay.**

□ The evidence does not suggest a linear relationship between the distance from the UC Berkeley campus and the price of a carne asada burrito in East Bay.

√ **If the slope of the true line is 0, it would be quite unlikely to have computed a confidence interval that did not capture 0.**

□ Given the confidence interval that we computed, the null hypothesis is false.

5 *Billboard* Hot 100 [13 points]

Rory is quite the music fan and noticed some interesting patterns in U.S. popular music from 2020 onwards. To take a closer look, she and the Data 8 staff compiled each of the songs that reached the top (position #1) of the *Billboard* Hot 100, the standard popularity chart for songs in the United States, in the table `billboard`. An excerpt is shown below.

- **Year** *(integer)*: The year that the song first reached position number 1 (#1) on the *Billboard* Hot 100 chart.
- **Artist** *(string)*: The primary artist of the song.
- **Name** *(string)*: The title of the song.
- **Genre** *(string)*: The main musical style of the song.
- **Weeks** *(integer)*: The number of weeks that the song has spent on the *Billboard* Hot 100 chart (at any position, 1 through 100), as of November 26, 2024. Songs may stay on the chart past the first year in which they charted.
- **Runtime** *(integer)*: The length of the album version of the song, in seconds.

Year	Artist	Name	Genre	Weeks	Runtime
2021	Lil Nas X	Industry Baby	Hip Hop	42	212
2024	Sabrina Carpenter	Please Please Please	Pop	24	186
2022	Taylor Swift	Anti-Hero	Pop	53	200
2020	BTS	Dynamite	Pop	32	199
2023	Morgan Wallen	Last Night	Country	60	163
2022	Future	Wait For U	R&B	41	190

...(73 rows omitted)

a. (3 points) Based on the data description, for which year is the distribution for songs first reaching #1 least likely to represent a typical distribution for songs first reaching #1 in any given year?

○ 2020 ○ 2021 ○ 2022 ○ 2023 √ **2024**

b. (4 points) To examine the distribution of the number of charting weeks for songs first reaching #1 in 2023, Richard created a histogram using `billboard.hist()`. The following table is the output he received from using `billboard.bin()`.

Bin	Count of songs
0	6
20	6
30	5
40	2
60	1
90	0

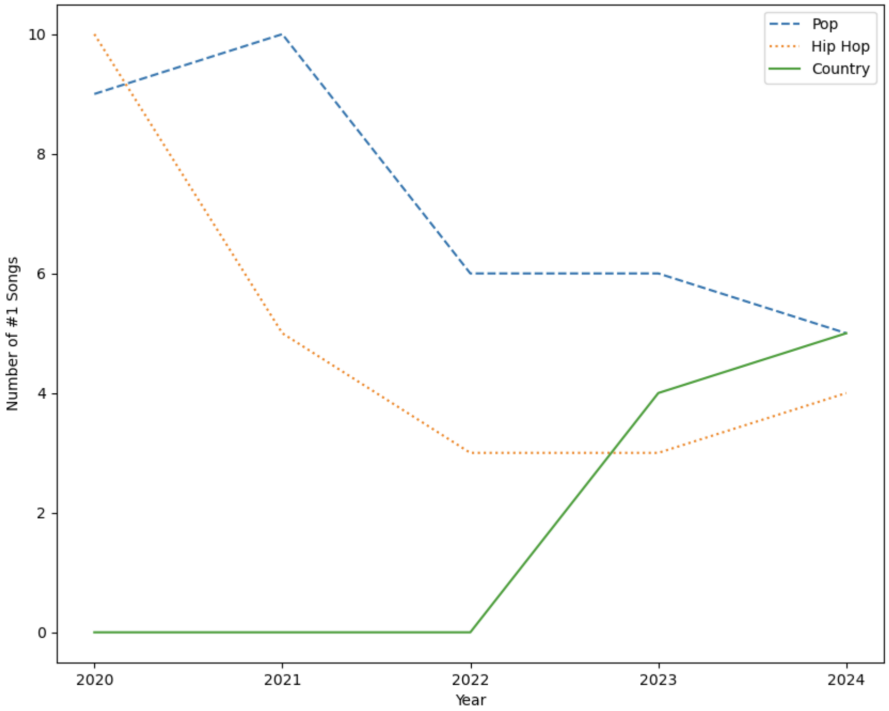
(i) (3 points) Based on the table, which do you expect to be smaller: the mean number of weeks that a 2023 song charted, or the median number of weeks that a 2023 song charted?

○ Mean √ **Median**

(ii) (1 point) How many songs have either charted for less than 20 weeks or have charted somewhere in the range of [40, 60] weeks? Give the correct number of songs or a range of possible numbers of songs.

○ 7
○ 8
√ **9**
○ 7 or 8
○ 9
○ 6 or 7

c. (6 points) Colin created the visualization below using the `billboard` table. Note: You do not need to worry about how the dashed, dotted, and straight lines were generated within this question part; this is just to help you distinguish the lines from one another.



(i) (2 points) True or False: The `where()` Table method was used in the creation of this plot.

√ **True** ○ False

(ii) (1 point) What is the dimension of the table directly before the plot is created, in rows and columns?

○ The dimensions of the `billboard` table
√ **5 rows and 4 columns**
○ 5 rows and 3 columns
○ 3 rows and 6 columns
○ 3 rows and 5 columns
○ 10 rows and 3 columns

(iii) (3 points) Choose the most *effective* and *appropriate* title for this plot.

○ "The amount of #1 country songs is increasing globally"
○ "#1 songs per year in the United States, by genre"
√ **"The genres of #1 songs on the Hot 100 have diversified in recent years"**
○ "Artists have shifted away from hip hop to country"

6 SpongeBob StatPants [14 points]

Andrew and a few other Data 8 staff members are big fans of the television show *SpongeBob SquarePants*. Each year, for the past three fall semesters, they have surveyed 100 enrolled students, for a total of 300 observations. Andrew compiles their survey data into the `spongebob` table. A six-row excerpt is shown below.

- **StudentID** *(integer)*: A code used to uniquely identify each student.
- **Age** *(float)*: The student's age on the first day of the fall semester, measured in years. This value may be a decimal.
- **Fall** *(integer)*: A categorical variable corresponding to the fall semester in which the student first enrolled at UC Berkeley.
- **TA** *(string)*: The student's Data 8 lab TA.
- **Character** *(integer)*: The number of *SpongeBob SquarePants* characters that the student was able to correctly identify when asked.

StudentID	Age	Fall	TA	Character
631	21	2024	Andrew	12
718	29	2022	Bing	2
733	21	2024	Ramisha	8
336	26	2024	Azalea	1
383	30	2024	Hailey	0
467	19	2023	Mia	10

...(294 rows omitted)

a. (4 points) Based on the data description, how many of the columns in `spongebob` contain numerical variables?

○ 1 √ **2** ○ 3 ○ 4 ○ 5

b. (3 points) Ramisha is interested in seeing the distributions of the 300 students' ages broken down by semester. Complete the **line of code** below to help Ramisha generate the proper visualization, making sure to include all relevant arguments.

Blank (a):

```
spongebob.___(a)___
```

Solution: `hist("Age", group = "Fall", unit = "Age")`

c. (3 points) Now, Kaed would like to see how the average character score has changed over the fall semesters for the 300 students.

First, help him generate a new table called `avg_scores` with exactly two columns, in the following order: one column containing the year of each unique fall semester in the `spongebob` table, and one column containing the average character scores for each of these semesters.

```
avg_scores = spongebob.___(b)___.___(c)___
```

(i) Fill in blank (b):

Solution: `select("Fall", "Character")`

(ii) Fill in blank (c):

Solution: `group("Fall", np.average)`

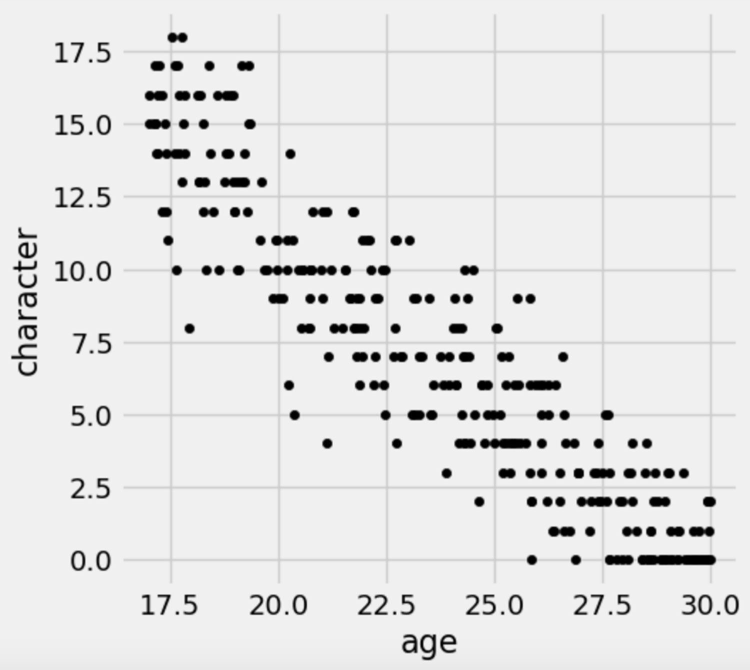
(iii) Then, complete the **line of code** in the `avg_scores` table to show how the character average has changed over time using an appropriate visualization.

```
avg_scores.___(d)___
```

Blank (d):

Solution: `plot("Fall", "Character average")`

d. (4 points) Simone generates a scatter plot using the **Age** and **Character variables**. Based on the scatter plot below, which of the following are valid conclusions that Simone can make? *Select all that apply.*



√ **The students who could name exactly 10 *SpongeBob* characters fall roughly in the age range of 17.5 to 25 years old.**

□ Being older in age causes students to correctly name fewer *SpongeBob* characters.

√ **Correctly naming more *SpongeBob* characters is associated with younger age.**

7 Bayesically Late with a Chance of Attendance Credit [7 points]

Aileen has a painfully early 11 AM class twice a week. Although, Aileen always makes it to class, there is only a 20% chance of making it on time. The class requires on-time attendance to be marked present, but Aileen's instructor, Conan, is nice and there is a 50% chance Aileen will be marked present even if late. Aileen's attendance habits are independent from day to day.

a. (2 points) A student is considered "chronically late" by the course staff if the probability that the student does not have perfect attendance (is marked present to both classes) in a given week is more than 50%. Is Aileen "chronically late"? **Show your work, including any calculations, in the box below.**

√ **Yes** ○ No

Solution:

$$P(\text{chronically late}) = 1 - P(\text{perfect attendance})$$

$$P(\text{perfect attendance}) = (0.2 \cdot 1 + 0.8 \cdot 0.5)^2$$

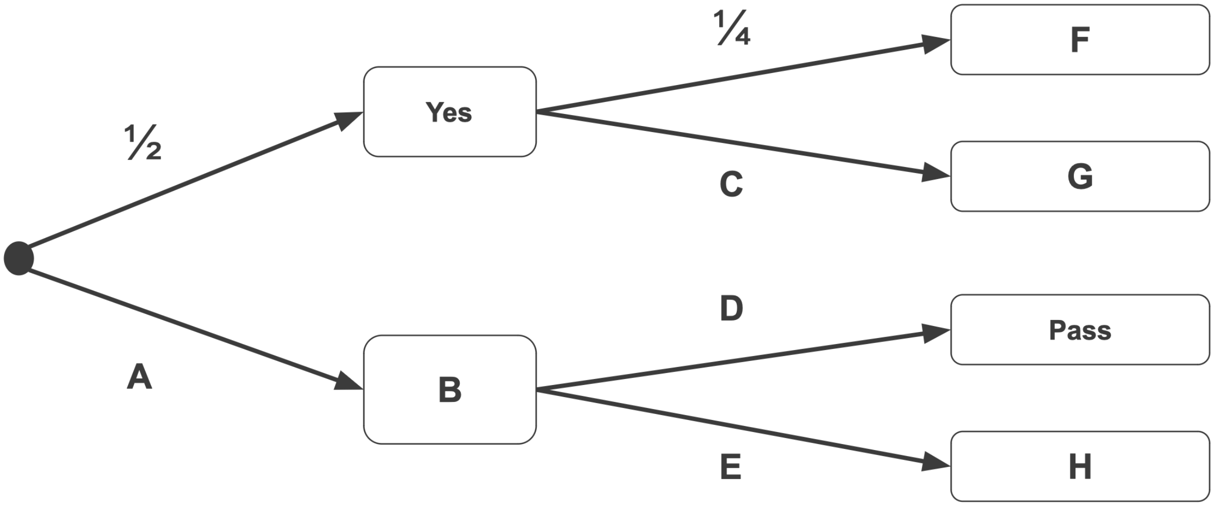
$$P(\text{perfect attendance}) = 0.36$$

$$P(\text{chronically late}) = 1 - P(\text{perfect attendance}) = 0.64$$

This probability is greater than 0.50, so we say that Aileen is chronically late.

b. (4 points) Conan has access to historical data from students who took the course before Aileen's semester, shown below. The cells in the pivot table below show the distribution of number of students who passed or failed the class, categorized by whether or not they were chronically late. Consider drawing a student at random from this historical data. **Using the boxes listed in parts (i) - (viii) below**, fill in the remainder of the tree diagram.

	Outcome	
Chronically Late	Pass	Fail
Yes	5	15
No	18	2



(i) Fill in letter A of the diagram:

Solution: $\frac{1}{2}$

(ii) Fill in letter B of the diagram:

Solution: No

(iii) Fill in letter C of the diagram:

Solution: $\frac{3}{4}$

(iv) Fill in letter D of the diagram:

Solution: $\frac{18}{20}$

(v) Fill in letter E of the diagram:

Solution: $\frac{2}{20}$

(vi) Fill in letter F of the diagram:

Solution: Pass

(vii) Fill in letter G of the diagram:

Solution: Fail

(viii) Fill in letter H of the diagram:

Solution: Fail

c. (1 point) Aileen passes the course. In the following semester, Sahand, a new student, takes the course and also passes. Assuming that Sahand is like a student drawn at random from historical data before his semester, what is the probability that he was chronically late? Write this probability in the box below.

Solution:

$$\frac{6}{24}$$

8 Congratulations! [0 points]

You have completed the Final. If you have not been told otherwise, you may bring all of your testing materials (reference sheet and this test paper), as well as your student ID, to the front of the room. Once you have been checked off, you may leave quietly.

- Please make sure that you have written your initials on each page of the exam. **You may lose points on pages where you have not done so.**
- Please make sure you have filled in circles and squares completely rather than having used a check mark, cross or any other mark.
- Double check that you have not skipped over any questions!

Below, you may draw and caption your favorite Data 8 experience or staff member!