Question #1

In a scientific study, 50 people suffering from insomnia were divided into two groups. A group of 20 subjects participated in a one-hour therapy session, while the other group consisting of the remaining 30 subjects did not receive any treatment. Three months later, 13 people in the therapy group reported improved sleep, while 12 people in the group not receiving therapy reported an improvement.

Part A: If these data were stored in "tidy" format with each case as a row and each variable as a column. How many rows and columns would the data frame contain? Do not consider any subject identifiers or variables not listed in the prompt. You do not need to explain your answer.

50 rows 2 columns

Part B: Of the variables present in this data set, identify which is the explanatory variable and which is the response variable. Briefly explain your answer using at most 2-sentences.

Explanatory - Treatment Response - Sleep Quality

Part C: Describe or sketch an appropriate data visualization that could be used to explore whether the explanatory and response variables you identified in Part B are associated. If providing a written description, limit your answer to no more than 2-sentences. If providing a sketch, you do not need to be overly precise so long as I can judge that it is the right type of graph.

Conditional bar plot

Part D: Create a contingency table summarizing the results of this study. Make sure to use the explanatory variable to define the table's rows and the response variable to define the table's columns.

Part E: Find the *odds ratio* that compares the odds of improved sleep in the group receiving therapy with the odds of improved sleep in the group not receiving therapy. Show your work for any calculations.

$$\frac{13 \cdot 18}{12 \cdot 7} = 2.78$$

Part F: Provide a 1-sentence interpretation of the odds ratio you found in Part E. Then, briefly indicate whether this odds ratio suggests an association between these variables. In total your entire response should be exactly 2-sentences.

odds of improved sleep 2.78 times higher in treatment.

As OR > 1 by a bit, suggests association b/w treat and sleep.

Question #2

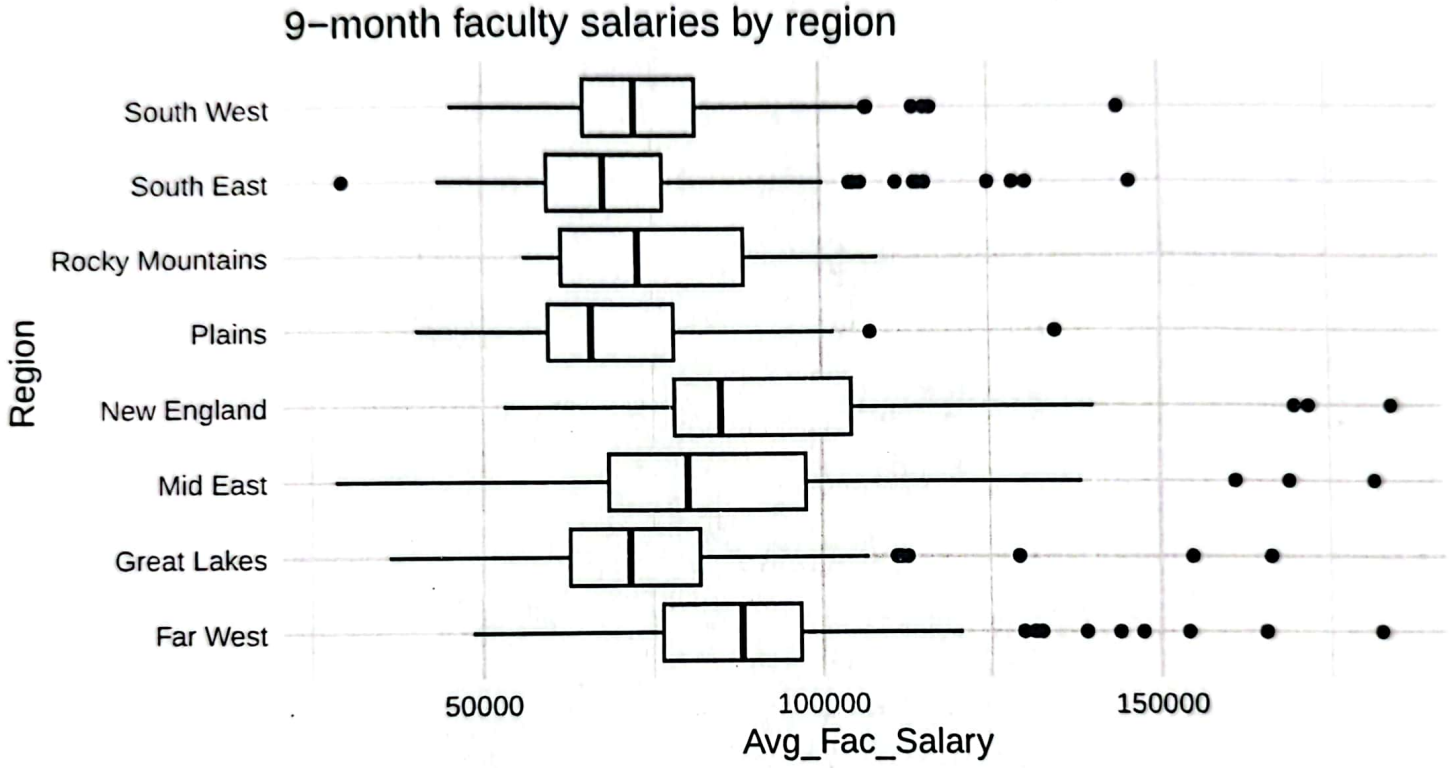
The data for this question are from the 2019 College Scorecard. We've previously used these data in class, though the data visualization and descriptive statistics below include only colleges with at least 1000 enrolled students.

1. "Region" - the census-designated geographic region where each college is located.
2. "Avg_Fac_Salary" - the average 9-month salary of the faculty members at each college.

		Improve Sleep: Yes	Improve Sleep: No
Treatment	Yes	13	7
	No	12	18

Table 1: Comparative summary of the median salaries of students from 1095 different colleges and universities according to geographic region

Region	N	Mean	StDev	Median	IQR
Far West	92	91289.15	23802.86	88060.5	20443.50
Great Lakes	156	74451.75	18011.65	71496.0	19140.75
Mid East	179	84800.92	22682.08	80037.0	29259.00
New England	65	92525.12	27103.22	84987.0	26325.00
Plains	97	69770.41	15275.53	65871.0	18513.00
Rocky Mountains	29	75944.48	15920.27	72819.0	27126.00
South East	238	70039.78	16814.71	67797.0	17118.00
South West	79	75846.30	17185.57	72468.0	16591.50



Part A: Is there an association between the variables "Region" and "Avg_Fac_Salary"? Explain your answer in at most 2-sentences.

Yes, large changes in mean/median b/w groups

Part B: Describe the distribution of "Avg_Fac_Salary" *within* the New England region. Limit your description to at most 2-sentences.

skewed right, median ≈ 85K

Part C: Using a *robust* descriptive statistic, which region exhibits the largest amount of variability in "Avg_Fac_Salary"? You do not need to explain your answer.

Robust ⇔ IQR Mideast

Part A: Using only the scatter plot of "ACT_median" vs. "Debt_median", qualitatively describe the relationship between these variables. Limit your response to at most 2-sentences.

Parabola shaped, loose association

Part B: Is it appropriate to rely upon Pearson's correlation coefficient to describe the relationship between "ACT_median" vs. "Debt_median"? Briefly explain, limiting your response to at most 2-sentences.

No, not linear

Part C: Interpret the effect of "ACT_median" in the the *simple linear regression* model where "ACT_median" is used to predict "Debt_median". Limit your response to a single sentence.

Each additional point for median ACT associated w/ $240 increase in debt.

Part D: Interpret the effect of "ACT_median" in the the *multivariable linear regression* model where "ACT_median" and "FourYearComp_Males" are used to predict "Debt_median".

Part E: Briefly explain why the estimated coefficient for "ACT_median" is positive in one model but negative in the other. How is this possible? And why does it happen? Limit your response to at most 4-sentences.

Part F: In Parts A and B you described the relationship between "ACT_median" and "Debt_median". Could your description be an example of the *ecological fallacy*? Briefly explain one way by which the ecological fallacy could occur in this situation. Limit your response to at most 3-sentences.