SDS/MTH 220: Badge Challenge 1

Name: _____

Instructions

- This badge challenge must be your own work entirely.
- This is an open mind, closed stats notebook, closed textbook, and closed fellow statistician badge challenge.
- All questions will be graded under the badge level grading system. On badge challenges, formatting and presentation guidelines are suspended, but you must show ***all*** work for computational answers and justify all claims for expository questions.
- Keep your explanations contextually meaningful and concise!
- You do not have to perform any long computations. For example, if the answer is 18.5, you will receive full credit for writing $2.5 * (4 + 3.5) - (1/2)^2$.
- Use the provided blank sheets of paper to write your answers. You may also use these pages for your scratch work.
- Please write on ***one side*** *of the blank pages* and staple any pages you want graded to the badge challenge. Your *answers should appear in* ***question order***.
- Put your name on the top right corner of each page that you submit and do not write where the staple will go.
- This is the first of four opportunities to demonstrate your current understanding of the first five badges for the course. You may complete as many badges as you would like. Questions left unattempted will not receive a badge level.

Name: _____

Question	Badge 1	Badge 2	Badge 3	Badge 4	Badge 5
1					
2					
3					
4					
5					
Recorded Score by Badge					

Question 1

Consider the Galton data, which contains the following variables among others:

`heights` - The child's height as an adult (inches)

`mother` - The child's mother's height as an adult (inches)

`sex` - The sex of the child (M or F)

Consider the following model:

```
heights = 43.15546 + 0.32655*mother + 1.96331*sexM + 0.05014*mother*sexM
```

(a) What is the reference level for this model? Justify your answer.

(b) Use the above model to construct a model formula between one's height and mother's height for females. Show or justify all work.

(c) Interpret the final coefficient in a contextually meaningful manner.

(d) Based on what you know about the above model, would you choose to keep the model as is or build a different trivariate model? Note that if you were to change the model, you must still use all of the three given variables and no others. Justify your choice.

Question 2

The Guild of Terrible Statisticians (GTS) has decided to investigate the importance and relevance of the Olympics to college students studying in the United States. The GTS work with Brown University to select the study's sample, by asking each varsity coach for a team roster. From each team roster, the GTS randomly samples 3 student-athletes to participate in the study. Each day of the Olympics, the participants are sent a text message asking if they watched Olympic coverage in the last 24 hours. They are then sent a quote from a member of Team USA and are asked on a scale of 1 to 7 how much the Olympics inspire their training. The GTS then collate the responses and report the daily average scores.

(a) Identify the population that GTS is interested in. State whether this study produces a representative sample of the population and give two reasons supporting your statement.

(b) Explain how this study falls victim to response bias **and** non-response bias.

(c) Do you expect the study's reported averages to be a good approximation for the population's opinion on the importance of the Olympics? Explain your answer.

(d) The GTS claim that the results from their study could be replicated using class lists instead of varsity team rosters. Do you agree? Justify your answer.

Question 3

You are investigating if countries get a 'bump' in their gold medal count when they host the Winter Olympics. Recalling that the 2010 Winter Olympics were in Vancouver Canada, you decide to use Canada as a first test case. You found the following Canadian gold medal tallies for the past 9 Winter Olympics:

Year	1984	1988	1992	1994	1998	2002	2006	2010	2014
Canada's Gold Medals	2	0	2	3	6	7	7	26	10

(a) What is the median number of gold medals that Team Canada have won over the past 9 Winter Olympics? Show all work.

(b) Find the 50% coverage interval. Show all work.

(c) Based on the above table and results, do you think that Canada got a 'bump' for being the host country? Justify your answer.

(d) You double check the table's data and notice that for the 2010 games, the listed number of medals is the total number and not the just the gold medals. The correct number of gold medals is actually 13 gold medals. Does the median number of gold medals that Team Canada has won over the past 9 Winter Olympics change? Does the mean change? Justify ***both*** your answers ***without*** doing any ***additional*** computations.

Question 4

Consider the models relating state's average Math SAT scores to state's average teacher salary and the state's fraction of students taking the exam, and their associated R^2 values:

Model	R^2 value
math $\sim$ frac	0.7558
math $\sim$ salary	0.1611
math $\sim$ frac + salary	0.7852

a) What is an R^2 value of a model? In your answer, be sure to explain how to compute it.

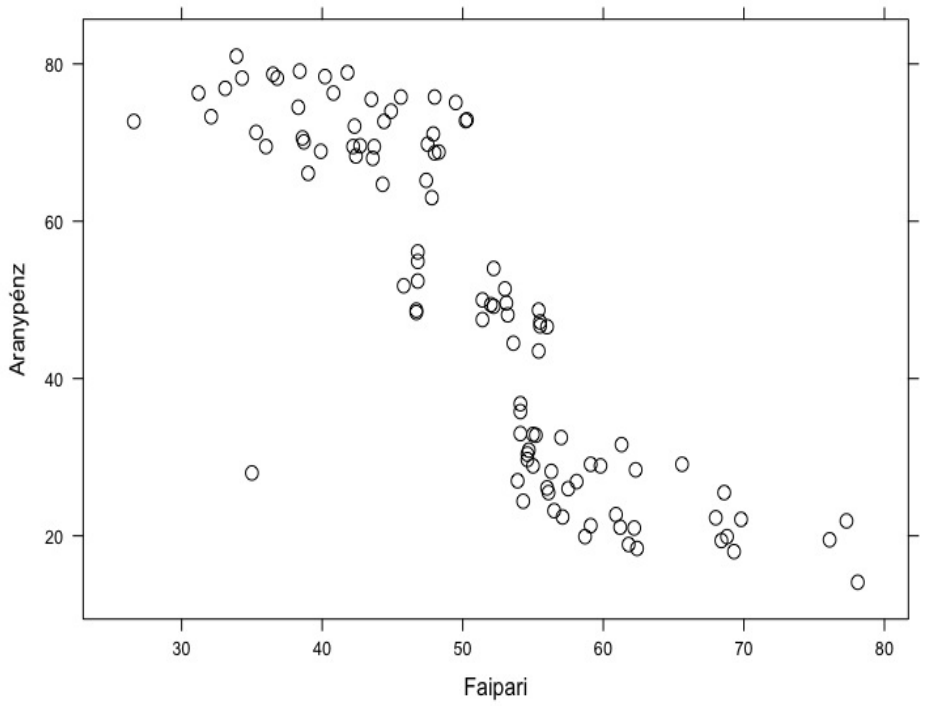
b) Compare the R^2 values of the first two models. Use your comparison to determine which of the explanatory variables is more effective at explaining the average math SAT scores.

c) Examining the R^2 values for all three models, what conclusion can you make about the efficacy of each explanatory variable on average math SAT scores? Justify your answer.

d) Which model do you think will have the highest R^2_{adj} value? Justify your answer.

Question 5

You are reading a blog post about the relationship between Aranypénz and Faipari. You are not an expert on these two variables or the domain where this data comes from. Relying on your statistical expertise, use the below plot to investigate these two variables.



(a) Find an approximate value for the mean for the Aranypénz variable and an approximate for the mean of the Faipari variable. Justify your approximations. Compare these values.

(b) You would like to compare the spread and skew for the Aranypénz and Faipari variables. Is this an appropriate plot for quickly determining the spread and skew of each variable? If not, what kind of visualization(s) would you recommend. Justify your answers.

(c) Bart Bivariate claims that there is an outlier in the dataset, while Uri Univariate says that neither variable has any outliers. Explain how they are both right. You may use images in your explanation.

(d) Given this above plot, what claims can you make about the relationship between these two variables? Be sure to support your claims.