STA 207 HW-1

Due Date: 9/3/2023 by 10:20AM (Submit in Moodle)

Derin Gezgin | Camel ID: 00468038

Problem-1 [38 points]

For each of the sets of variables, identify the response variable (Y) and independent variables (X) in the study. Then, identify if the variables are quantitative (numerical: continuous or discrete) or qualitative (categorical: ordinal or nominal).

a.) [12 points] A study is conducted to understand how the state average SAT Math scores vary based on the average SAT Verbal scores. The variables are SAT Math scores, SAT verbal scores, percentage of eligible students taking the SAT, percentage of adult population without a high school education, average annual teacher salary, and state population.

Response Variable (Y):

I. **SAT Math Scores** ***(Numerical – Discrete)***
It's *numerical* because it's measured in a numerical value. Its sub-category is *discrete* because we can list all possible math scores from 0 to 800 and there's finite number of possibilities.

Independent Variables (X):

I. **SAT Verbal scores (*Numerical – Discrete)***
Same as Math scores, it's measured in a *numerical* value. Its sub-category is also *discrete* because there's finite number of possibilities.

II. **If a student in the population taking the SAT *(Categorical – Nominal)***
This is a *categorical* variable because it's measured in a Yes/No format. Yes/No questions don't have an order so it's *nominal*. On the other hand, in the study it's used in a numerical variable as a percentage.

III. **If an adult in the population has high school education. *(Categorical – Nominal)***
Like the previous variable, this is a *categorical* variable because it's measured in a Yes/No format which doesn't have an order so it's *nominal*. On the other hand, in the study it's used in a numerical variable as a percentage.

IV. **Annual teacher salary. *(Numerical – Discrete)***
It's a *numerical* variable because it's measured as a numerical value. It's *discrete* because in our current system currency is limited up to 2 decimal points. Depending on how use it, it can also be interpreted as *continuous*.

V. **State population *(Numerical – Discrete)***
This is a *numerical* variable because it's measured as a numerical value. It's *discrete* because we can't have a population value in decimals.

b.) [12 points] To see how shooting percentage, average total points, rebounds, assists, and turnover percentage affect a team's winning percentage.

Response Variable (Y):

I. **Winning count of the team *(Numerical – Discrete*)***
As it's measured as a numerical value and it has finite number of possible answers, it's a *numerical discrete* variable. But in the study, it's used in the percentage format which makes it *continuous* if we consider the format it's used.

Independent Variables (X):

I. **Number of Shoots** ***(Numerical – Discrete*)***
As it's measured as a numerical value, it's a *numerical* variable. Number of shoots can't have decimals so it's continuous but, in the study, it's used as a percentage which would make it continuous if we consider the way it's used.

II. **Total points** ***(Numerical – Discrete)***
As it's a numerical value which has a finite number of possible values it's a *numerical discrete* variable.

III. **Total rebounds** ***(Numerical – Discrete)***
Like the "total points," as it's a numerical value with finite number of possible values, it's a *numerical discrete* variable.

IV. **Total assists** ***(Numerical – Discrete)***
Because of the same reason with the previous 2 variables, it's also *numerical discrete.*

V. **Turnover Count** ***(Numerical – Discrete*)***
If we consider how it's measured and it being a numerical value, this is a *numerical* and *discrete* variable. But in the study, it's used as a percentage which would make it *continuous* if we consider how it's used in the study.

c.) [14 points] To evaluate the effect of lifestyle factors on sleep efficiency, an experiment is conducted and variables including caffeine consumption, alcohol consumption, smoking status, exercise frequency is the # of times exercise in a week, age, gender, and duration of sleep are recorded.

Response Variable (Y):

I. **Duration of Sleep** ***(Numerical – Continuous*)***
It's *numerical* because it's measured as a numerical value. It's *continuous* because time measurement can have a high precision. It can also be interpreted as *discrete* depending on how it's measured. If it's measured as whole hour format (1…24) it'd be discrete. It's used to assess the sleeping efficiency.

Independent Variables (X):

II. **Caffeine consumption** ***(Numerical – Continuous)***
This is a *numerical* variable as it's measured as a numerical value. It's *continuous* as it's measured in a scale with high precision.

III. **Alcohol consumption** ***(Numerical – Continuous)***
This is the same case as the previous variable, caffeine consumption.

IV. **Smoking status** ***(Categorical – Nominal)***
As this is a variable that can have Yes/No value, it's *categorical*. Yes/No doesn't have a natural order so it's *nominal*.

V. **Exercise frequency (# of times exercise in a week)** ***(Numerical – Discrete)***
This is a *numerical* variable because it's measured as a numerical value. It's *discrete* as there's finite number of possibilities without any decimals.

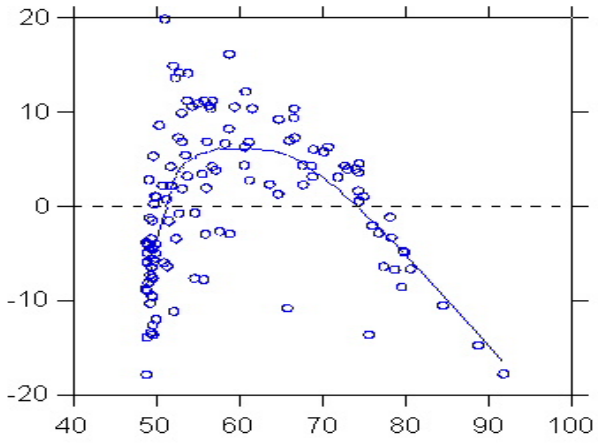
VI. **Age** ***(Numerical – Continuous*)***
This is a *numerical* variable because it's measured as a numerical value. Depending on how it's measured in the study it can be both *discrete* and continuous. If it's measured as whole numbers, it's discrete, if it's measured as decimal numbers with higher accuracy, it'd be categorized as *continuous*.

VII. Gender *(Categorical – Nominal)*

Gender is a *categorical* as it's measured in a non-numerical way. As gender doesn't have a natural order it's *nominal*.

Problem-2 [5 points]

Discuss what might be the value of the correlation between the X and Y variable using this graph. Explain why.



The correlation between the X and the Y is approximately 0 because the graph is not a linear graph but it's a curve. As it's a curve, it's not possible to assess a linear relation.

Problem-3 [7 points]

Share an interesting study (with reference to the article/blog) where correlation is used inaccurately to imply causation and write a summary.

Example 1:

https://obamawhitehouse.archives.gov/the-press-office/2015/10/01/statement-president-shootings-umpqua-community-college-roseburg-oregon

In this speech made by Barrack Obama -44th president of the USA- he made the following statement:

"We know that states with the most gun laws tend to have the fewest gun deaths. So the notion that gun laws don't work, or just will make it harder for law-abiding citizens and criminals will still get their guns is not borne out by the evidence."

In this statement, Obama claims, the fact that states with the most gun laws tend to have the fewest gun deaths (**correlation**), means that stricter gun laws are effective in reducing gun deaths (**causation**).

However, as we said before, *correlation does not imply causation*, and while there's a relationship between the amount of gun laws and gun related deaths, we can't have a

conclusion of gun laws are the direct cause of the reduction in deaths. There might be compounding variable like law enforcement presence, culture in the state, economic conditions that can also affect this situation.

I found this example while looking for different articles in *factcheck.org*. I wanted to mention this.

https://www.factcheck.org/2015/10/gun-laws-deaths-and-crimes/

Another Website I Found:

https://www.tylervigen.com/spurious-correlations

In this website, there's more than 3,000 examples for data pairs showing correlation (for example UFO sightings in South Carolina and total number of successful Everest climbs) but of course any of these data pairs imply causation. Also, this website generates a research paper using a LLM and share it.