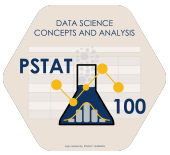


INFORMATION ABOUT IN-CLASS ASSESSMENT 01

PSTAT 100: Data Science Concepts and Analysis
Spring 2024, with Ethan P. Marzban

SOLUTIONS TO SOME SAMPLE PRACTICE PROBLEMS

Questions 1 - 4 refer to the following situation: Morgana wants to know whether or not UCSB students find studying in the library to be beneficial to their education (as opposed to studying somewhere else). To that end, she decides to conduct a study.

In order to ensure a somewhat diverse representation of students, Morgana decides to randomly select 5 lectures taking place on campus right now. From each of these 5 lectures, she takes a simple random sample of 10 students and administers them an email survey asking the following two questions:

1. Do you regularly study in the library?
2. What score did you receive on your most recent exam?

She then compares the exam scores of the students who regularly study in the library to those who do not.

1) Did Morgana conduct an observational study or an experiment?

A) Observational Study
B) Experiment

2) Which sampling technique did Morgana utilize to collect her sample of 50 participants?

A) Simple random sampling
B) Stratified random sampling
C) Cluster sampling

3) Which of the following best describes (in words) the target population of Morgana's study?

A) All UCSB students
B) UCSB students who regularly study in the library
C) UCSB students who regularly study off-campus
D) All college students in the uS
E) None of the above

4) ~~Since Morgana is conducting an email survey, we can consider the access frame of Morgana's study to be "the set of all people who attend lectures." Which of the following (fictional) individuals would be included in neither the target population nor the access frame?~~

~~A) John, a UCSB student who does not regularly attend lecture~~
~~B) Jack, a UCSB student who regularly attends lecture~~
~~C) Jane, a UCLA student who does not regularly attend lecture~~
~~D) Jill, a UCLA student who does not regularly attend lecture~~
~~E) None of the above~~

Please disregard this question - it was poorly worded. The main point of this problem was to illustrate that there are certain individuals who lie neither in the target population nor the access frame. Here, we can consider the access frame to simply be the set of people who attend lecture and are reachable by email; hence someone who doesn't attend lecture and doesn't check their email, and is also **not** a UCSB student, would be neither a part of the access frame nor the target population.

<u>Questions 5 - 9 refer to the following situation:</u> Consider the relations X and Y, depicted below:

X

Food	Type
Apple	Fruit
Kale	Vegetable
Banana	Fruit
Spinach	Vegetable

Y

Food	Color
Apple	Red
Kale	Green
Guava	Pink

Table I

Food	Type	Color
Apple	Fruit	Red
Kale	Vegetable	Green
Banana	Fruit	NA
Spinach	Vegetable	NA

Table II

Food	Type	Color
Apple	Fruit	Red
Kale	Vegetable	Green

Table III

Food	Type	Color
Apple	Fruit	Red
Kale	Vegetable	Green
Banana	Fruit	NA
Spinach	Vegetable	NA
Guava	NA	Pink

Table IV

Food	Type	Color
Apple	Fruit	Red
Kale	Vegetable	Green
Guava	NA	Pink

5) Which table is the result of performing an inner join of Y onto X (i.e. `inner_join(X, Y)`, assuming all join keys are specified correctly)?

(A) Table I (B) Table II (C) Table III (D) Table IV
(E) None of the above

6) Which table is the result of performing a left join of Y onto X (i.e. `left_join(X, Y)`, assuming all join keys are specified correctly)?

(A) Table I (B) Table II (C) Table III (D) Table IV
(E) None of the above

7) Which table is the result of performing a full join of Y onto X (i.e. `full_join(X, Y)`, assuming all join keys are specified correctly)?

(A) Table I (B) Table II (C) Table III (D) Table IV
(E) None of the above

8) Which table is the result of performing a right join of Y onto X (i.e. `right_join(X, Y)`, assuming all join keys are specified correctly)?

(A) Table I (B) Table II (C) Table III (D) Table IV
(E) None of the above

9) Which of the following could be the primary key of the X relation?

(A) {**Food**}
(B) {**Type**}
(C) {**Food, Type**}
(D) Such a primary key does not exist.
(E) None of the above.

Questions 10 - 13 refer to the following situation: Consider the following code, which is designed to create a small dataframe containing information about pets at a local veterinarian hospital:

```
library(tidyverse)

pet_roster <- data.frame(
    Name = c("Mr. Fluffy",   "Captain Barkles",   "Steve"),

    Species = c("Cat",   "Dog",   "Cat"),

    Age = c(3, 1, NA),

    Weight = c(6, 8, 4),
)
```

10) What will be the result of running the command
> **`nrow(pet_roster)`**?

(A) 1 (B) 2 (C) 3 (D) 4 (E) NA

11) What will be the result of running the command
> **`nrow(names(pet_roster))`**?

(A) 1 (B) 2 (C) 3 (D) 4 (E) NULL

12) Which of the following commands would correctly produce a table displaying the average (median) weight of each species of pet?

(A)
```
pet_roster %>%
    summarise(Med_Weight = median(Weight))
```

(B)
```
pet_roster %>%
    summarise(Med_Weight = median(Weight)) %>%
    group_by(Species)
```

(C)
```
pet_roster %>%
    group_by(Species) %>%
    summarise(Med_Weight = median(Weight))
```

(D) None of the above

13) It makes sense to encode the Species variable as an unordered factor. Which of the following commands will, when run, convert the Species variable to an unordered factor while still keeping the variable names the same, and assign this modified dataframe to a variable called `pet_roster_mod`?

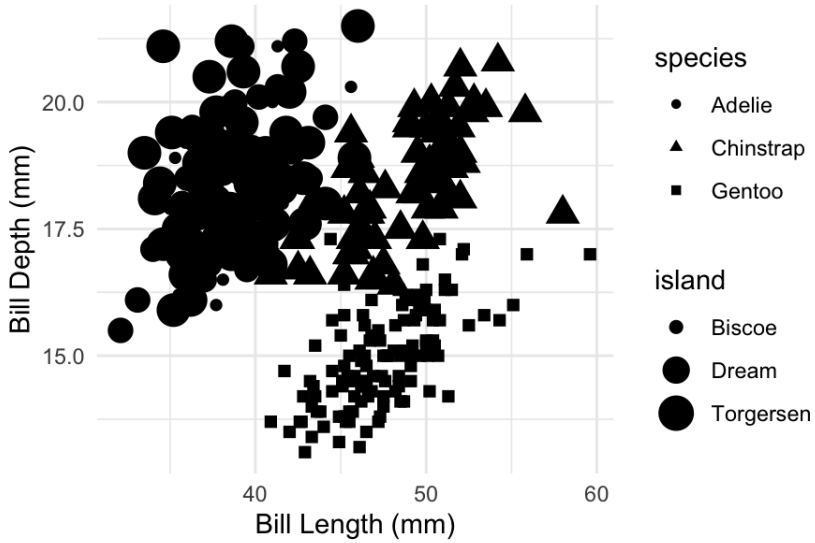
(A)
```
pet_roster %>%
    mutate(Species = factor(Species))
```

(B)
```
pet_roster_mod <- pet_roster %>%
    mutate(Species = factor(Species))
```

(C)
```
pet_roster_mod %>% factor(pet_roster)
```

(D) None of the above

Questions 14 - 17 refer to the following plot: (plot generated using the `palmerpenguins` dataset)



14) How many variables have information encoded in the figure above?

(A) 3 (B) 4 (C) 5 (D) 6 (E) 7

Remember that we, for this class, treat x- and y- coordinates as separate aesthetics.

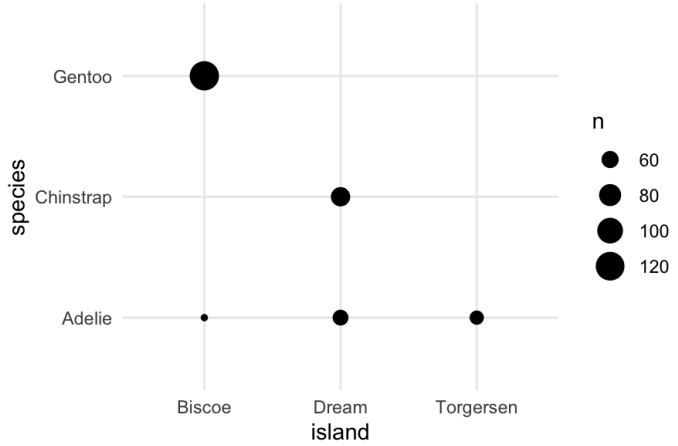
15) How many aesthetics in the figure above are being modified to encode information stored in a variable?

(A) 3 (B) 4 (C) 5 (D) 6 (E) 7

16) If we instead wanted to use color to encode information encoded in the species variable, which color scale should we use?

(A) Qualitative (B) Sequential (C) Diverging

17) Consider the following plot:



What is this type of plot called?

(A) Barplot (B) Balloon plot (C) Side-by-side boxplot
(D) None of the above

---

18) How would we go from table I (below) to table II?

Table I

City	Type	Value
Santa Barbara	2020_pop	88,695
Santa Barbara	2010_pop	88,544
San Francisco	2020_pop	870,014
San Francisco	2010_pop	805,519

Table II

City	2020_pop	2020_pop
Santa Barbara	88,695	88,544
San Francisco	870,014	805,519

A) Just **melt** using Type as the colvar and extracting values from the Value column
B) Just **pivot**, getting column names from the Type column and values from the Value column
C) Melt and then pivot, in sequence
D) None of the above