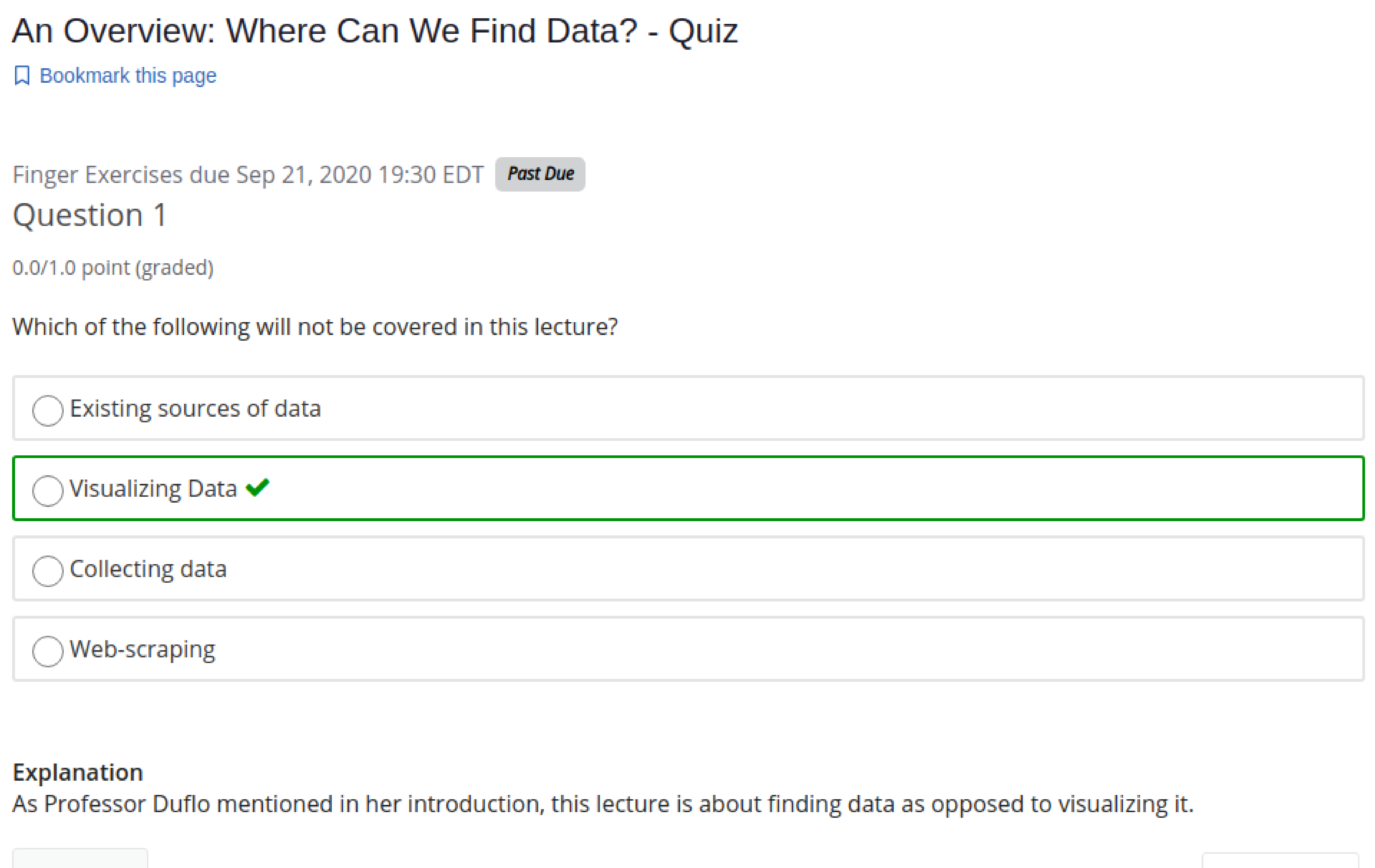


An Overview: Where Can We Find Data? - Quiz

Finger Exercises due Sep 21, 2020 19:30 EDT **Past Due**

Question 1

0.0/1.0 point (graded)

Which of the following will not be covered in this lecture?

○ Existing sources of data

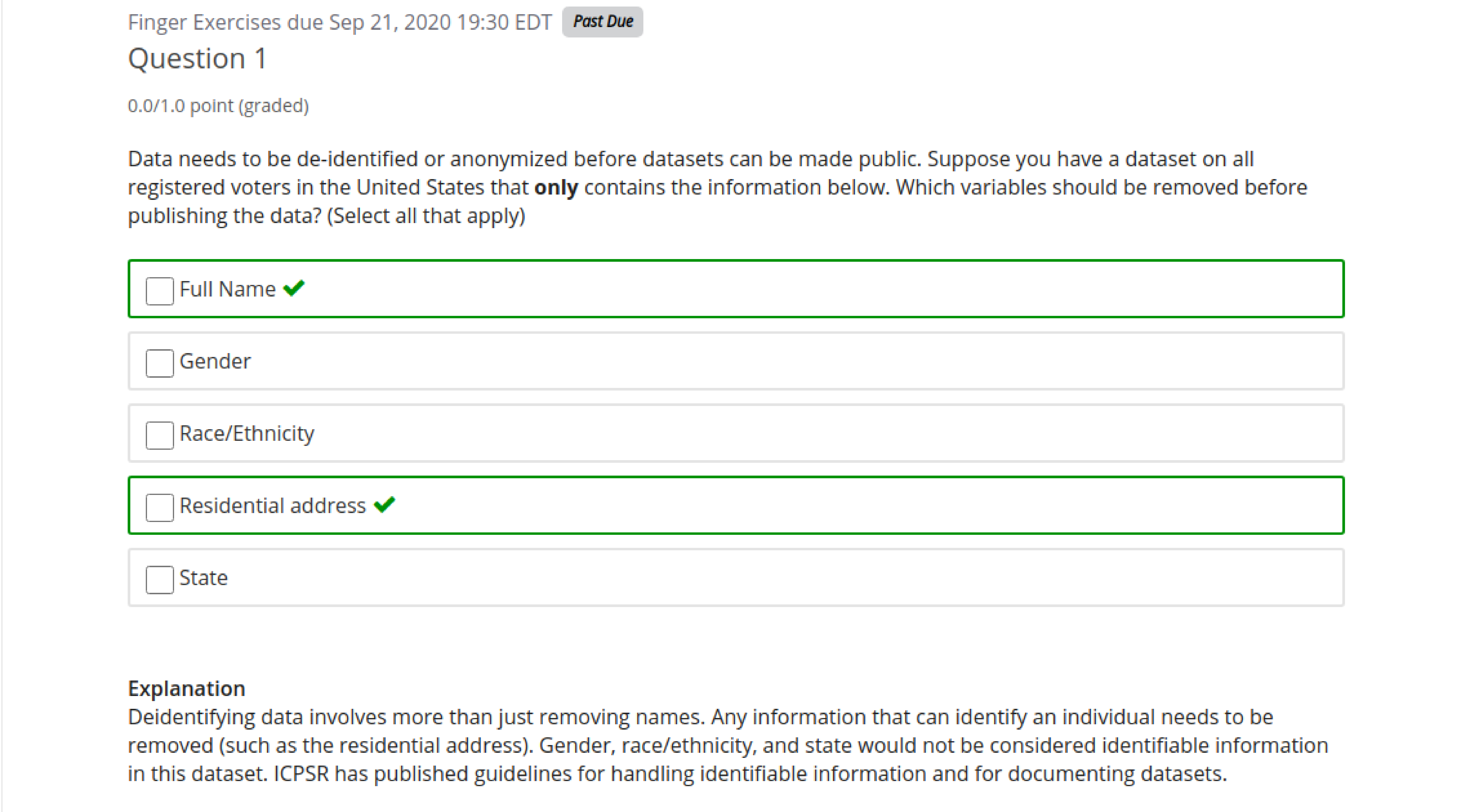
○ Visualizing Data ✓

○ Collecting data

○ Web-scraping

Explanation

As Professor Duflo mentioned in her introduction, this lecture is about finding data as opposed to visualizing it.



Finger Exercises due Sep 21, 2020 19:30 EDT **Past Due**

Question 1

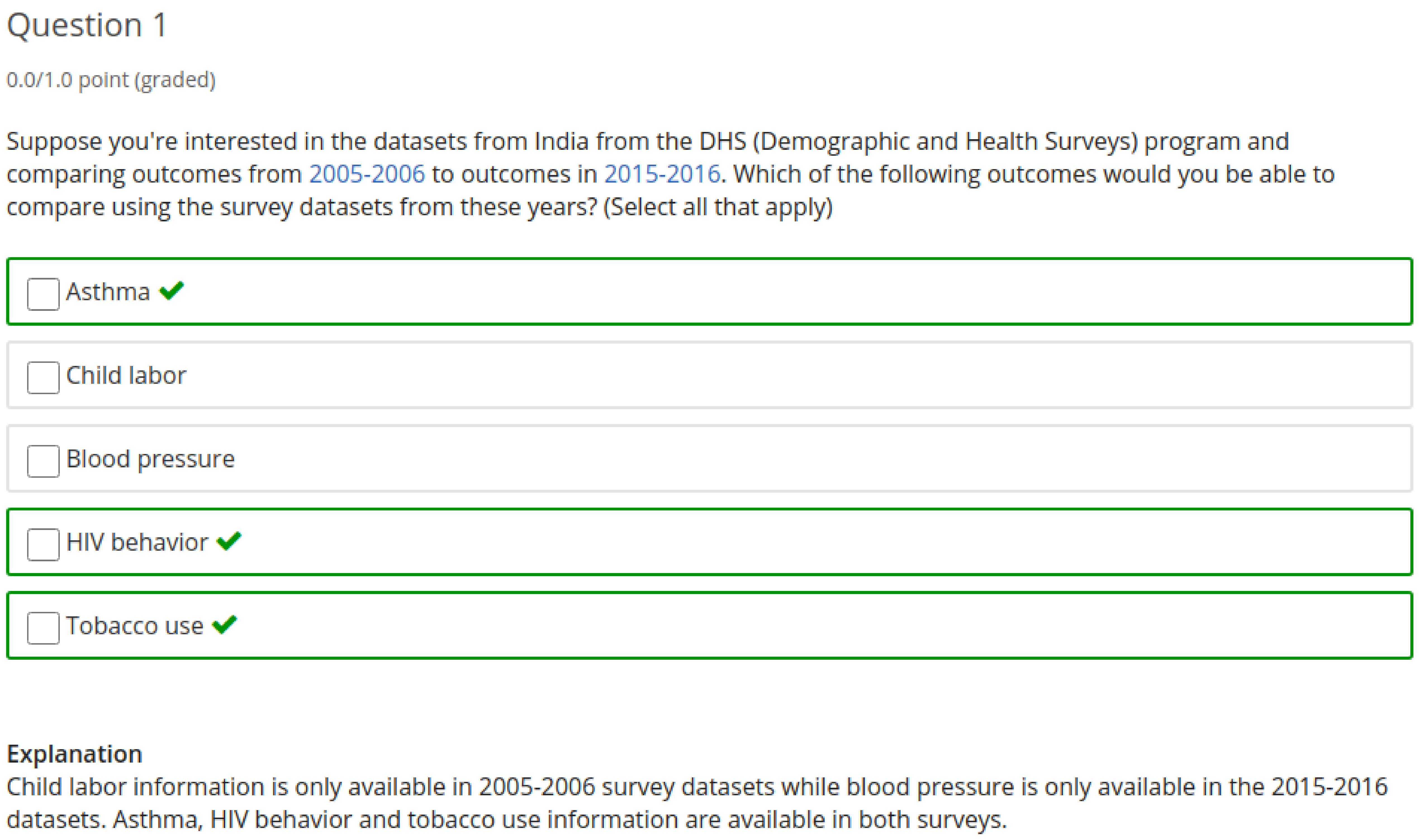
0.0/1.0 point (graded)

Data needs to be de-identified or anonymized before datasets can be made public. Suppose you have a dataset on all registered voters in the United States that **only** contains the information below. Which variables should be removed before publishing the data? (Select all that apply)

- □ Full Name ✓
- □ Gender
- □ Race/Ethnicity
- □ Residential address ✓
- □ State

Explanation

Deidentifying data involves more than just removing names. Any information that can identify an individual needs to be removed (such as the residential address). Gender, race/ethnicity, and state would not be considered identifiable information in this dataset. ICPSR has published guidelines for handling identifiable information and for documenting datasets.



Question 1

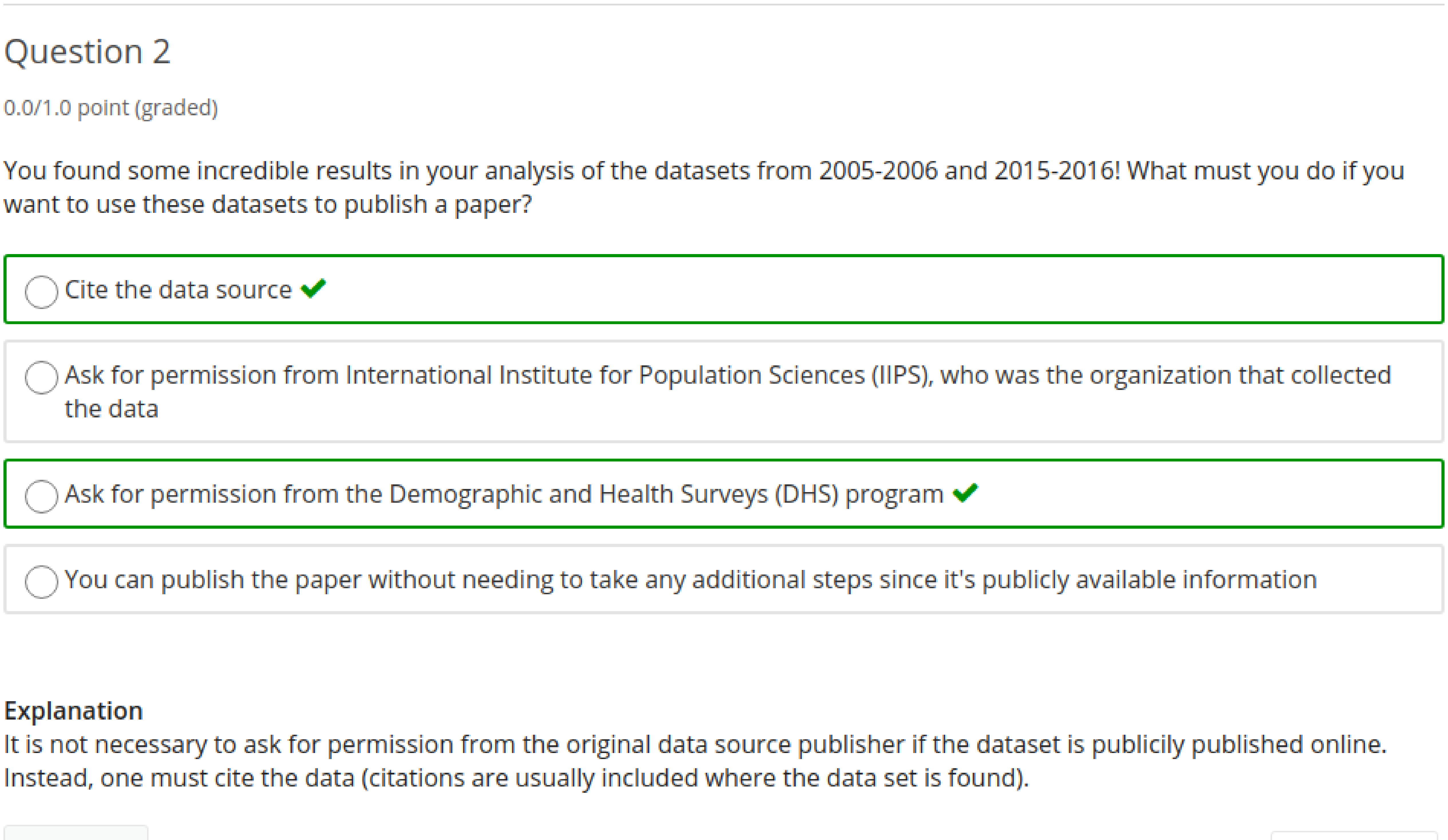
0.0/1.0 point (graded)

Suppose you're interested in the datasets from India from the DHS (Demographic and Health Surveys) program and comparing outcomes from 2005-2006 to outcomes in 2015-2016. Which of the following outcomes would you be able to compare using the survey datasets from these years? (Select all that apply)

- ☐ Asthma ✓
- ☐ Child labor
- ☐ Blood pressure
- ☐ HIV behavior ✓
- ☐ Tobacco use ✓

Explanation

Child labor information is only available in 2005-2006 survey datasets while blood pressure is only available in the 2015-2016 datasets. Asthma, HIV behavior and tobacco use information are available in both surveys.



Question 2

0.0/1.0 point (graded)

You found some incredible results in your analysis of the datasets from 2005-2006 and 2015-2016! What must you do if you want to use these datasets to publish a paper?

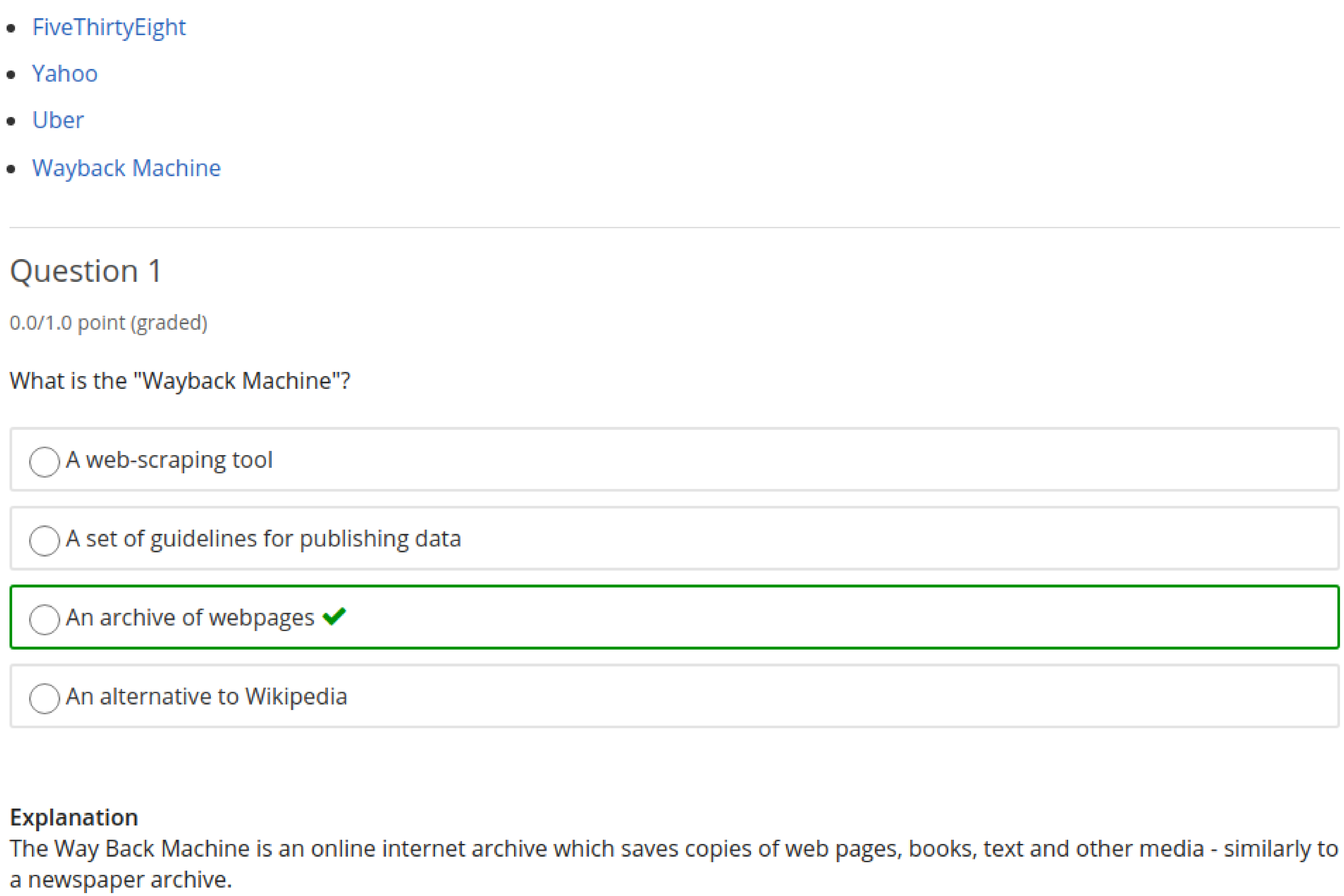
○ Cite the data source ✓

○ Ask for permission from International Institute for Population Sciences (IIPS), who was the organization that collected the data

○ Ask for permission from the Demographic and Health Surveys (DHS) program ✓

○ You can publish the paper without needing to take any additional steps since it's publicly available information

Explanation
It is not necessary to ask for permission from the original data source publisher if the dataset is publicily published online. Instead, one must cite the data (citations are usually included where the data set is found).



- FiveThirtyEight
- Yahoo
- Uber
- Wayback Machine

Question 1

0.0/1.0 point (graded)

What is the "Wayback Machine"?

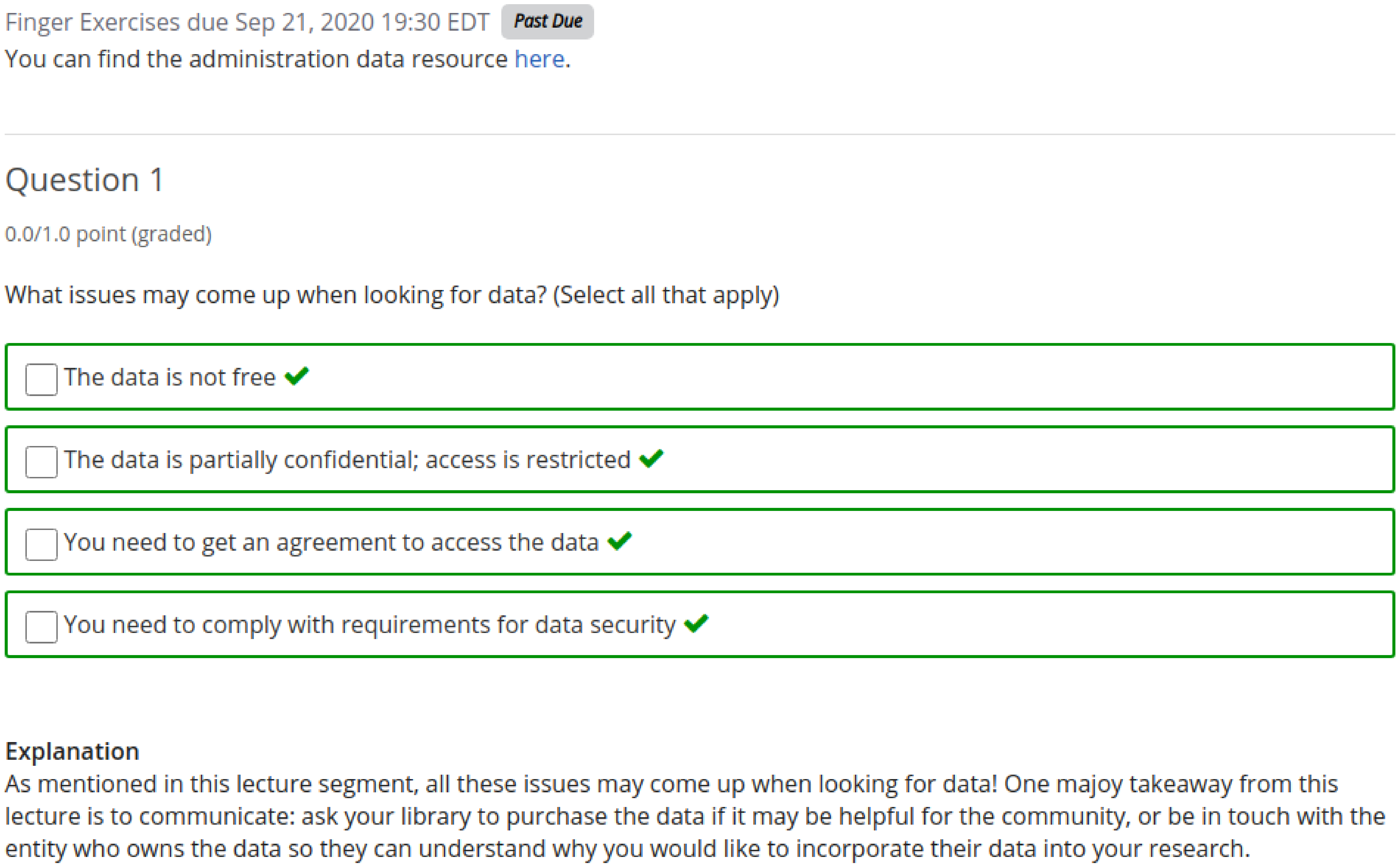
○ A web-scraping tool

○ A set of guidelines for publishing data

○ An archive of webpages ✓

○ An alternative to Wikipedia

Explanation
The Way Back Machine is an online internet archive which saves copies of web pages, books, text and other media - similarly to a newspaper archive.



Finger Exercises due Sep 21, 2020 19:30 EDT **Past Due**

You can find the administration data resource here.

Question 1

0.0/1.0 point (graded)

What issues may come up when looking for data? (Select all that apply)

□ The data is not free ✓

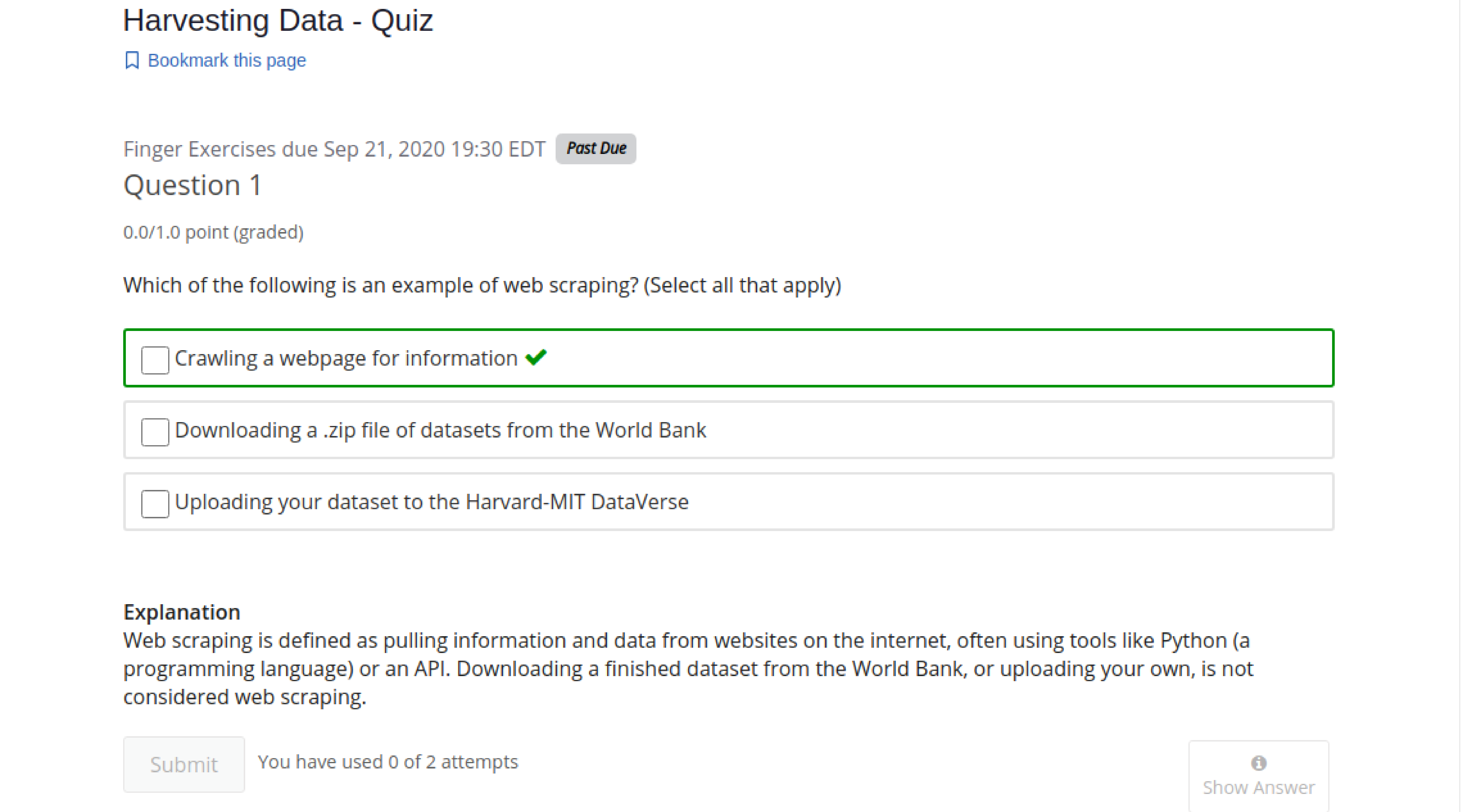
□ The data is partially confidential; access is restricted ✓

□ You need to get an agreement to access the data ✓

□ You need to comply with requirements for data security ✓

Explanation

As mentioned in this lecture segment, all these issues may come up when looking for data! One majoy takeaway from this lecture is to communicate: ask your library to purchase the data if it may be helpful for the community, or be in touch with the entity who owns the data so they can understand why you would like to incorporate their data into your research.



Harvesting Data - Quiz

Finger Exercises due Sep 21, 2020 19:30 EDT *Past Due*

Question 1

0.0/1.0 point (graded)

Which of the following is an example of web scraping? (Select all that apply)

☐ Crawling a webpage for information ✓

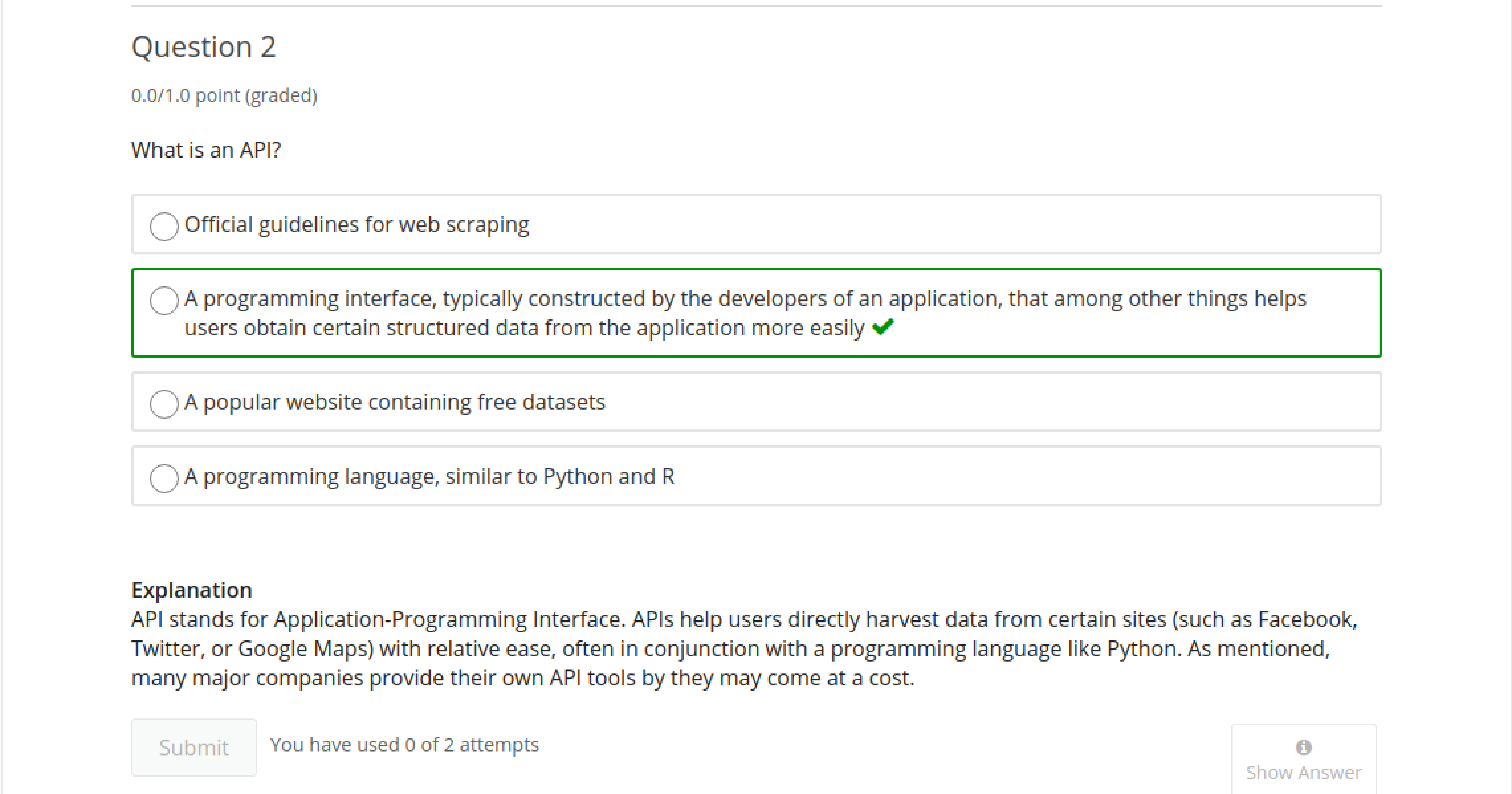
☐ Downloading a .zip file of datasets from the World Bank

☐ Uploading your dataset to the Harvard-MIT DataVerse

Explanation

Web scraping is defined as pulling information and data from websites on the internet, often using tools like Python (a programming language) or an API. Downloading a finished dataset from the World Bank, or uploading your own, is not considered web scraping.

You have used 0 of 2 attempts



Question 2

0.0/1.0 point (graded)

What is an API?

○ Official guidelines for web scraping

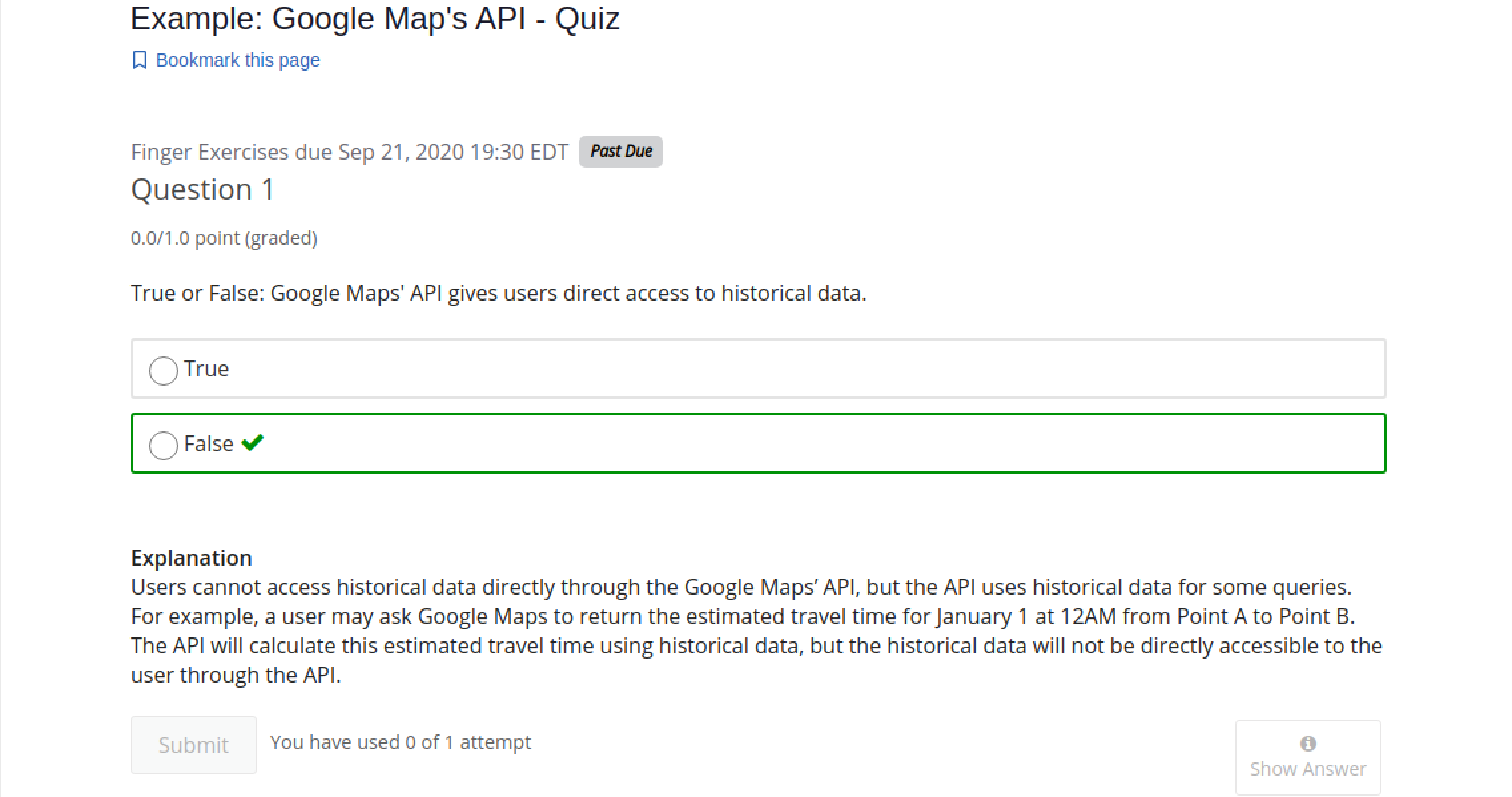
○ A programming interface, typically constructed by the developers of an application, that among other things helps users obtain certain structured data from the application more easily ✓

○ A popular website containing free datasets

○ A programming language, similar to Python and R

Explanation
API stands for Application-Programming Interface. APIs help users directly harvest data from certain sites (such as Facebook, Twitter, or Google Maps) with relative ease, often in conjunction with a programming language like Python. As mentioned, many major companies provide their own API tools by they may come at a cost.

You have used 0 of 2 attempts



Example: Google Map's API - Quiz

Bookmark this page

Finger Exercises due Sep 21, 2020 19:30 EDT *Past Due*

Question 1

0.0/1.0 point (graded)

True or False: Google Maps' API gives users direct access to historical data.

○ True

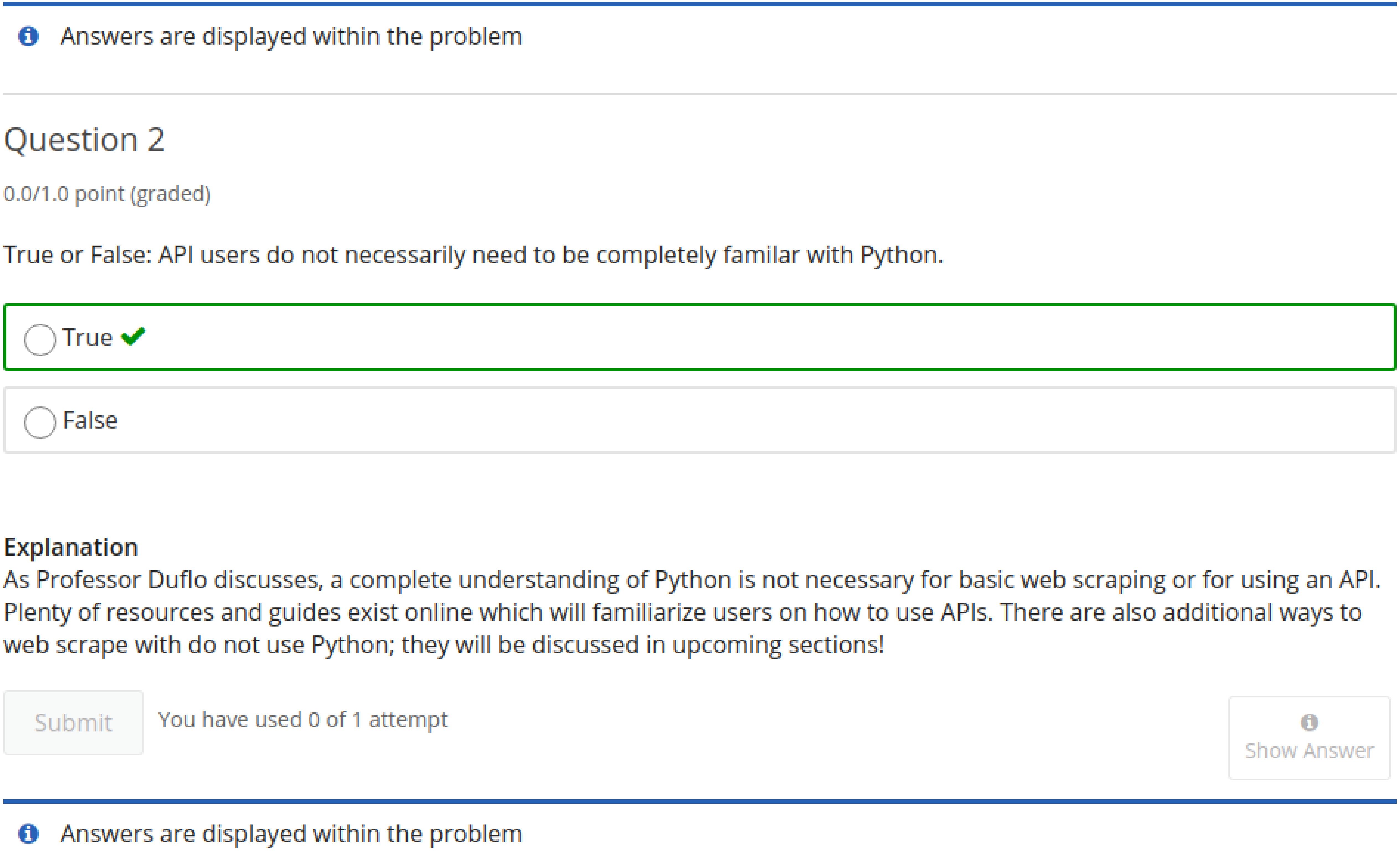
○ False ✓

Explanation

Users cannot access historical data directly through the Google Maps' API, but the API uses historical data for some queries. For example, a user may ask Google Maps to return the estimated travel time for January 1 at 12AM from Point A to Point B. The API will calculate this estimated travel time using historical data, but the historical data will not be directly accessible to the user through the API.

Submit You have used 0 of 1 attempt

Show Answer



ℹ Answers are displayed within the problem

Question 2

0.0/1.0 point (graded)

True or False: API users do not necessarily need to be completely familar with Python.

○ True ✓

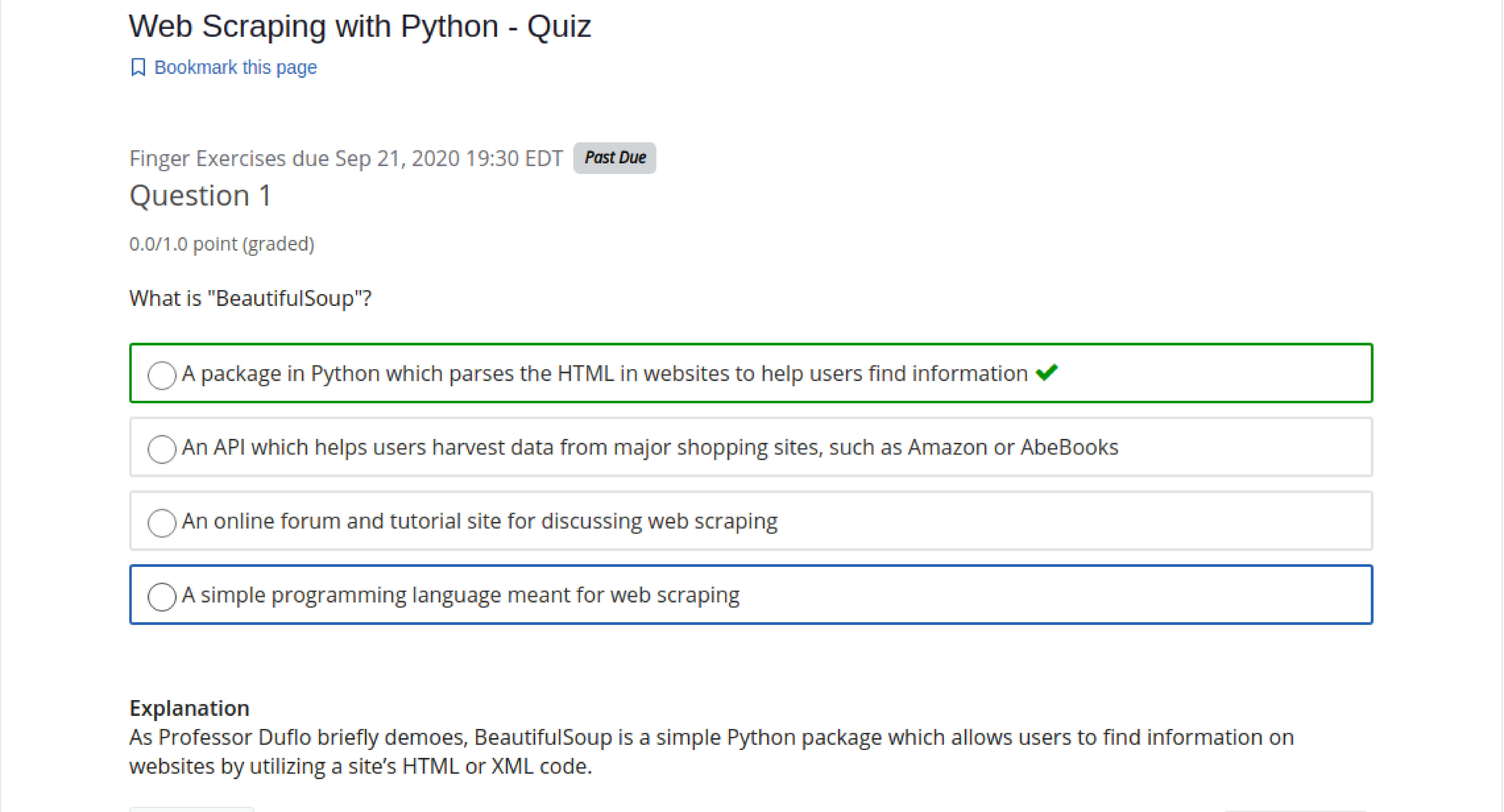
○ False

Explanation

As Professor Duflo discusses, a complete understanding of Python is not necessary for basic web scraping or for using an API. Plenty of resources and guides exist online which will familiarize users on how to use APIs. There are also additional ways to web scrape with do not use Python; they will be discussed in upcoming sections!

You have used 0 of 1 attempt

ℹ Answers are displayed within the problem



Web Scraping with Python - Quiz

Finger Exercises due Sep 21, 2020 19:30 EDT *Past Due*

Question 1

0.0/1.0 point (graded)

What is "BeautifulSoup"?

○ A package in Python which parses the HTML in websites to help users find information ✓

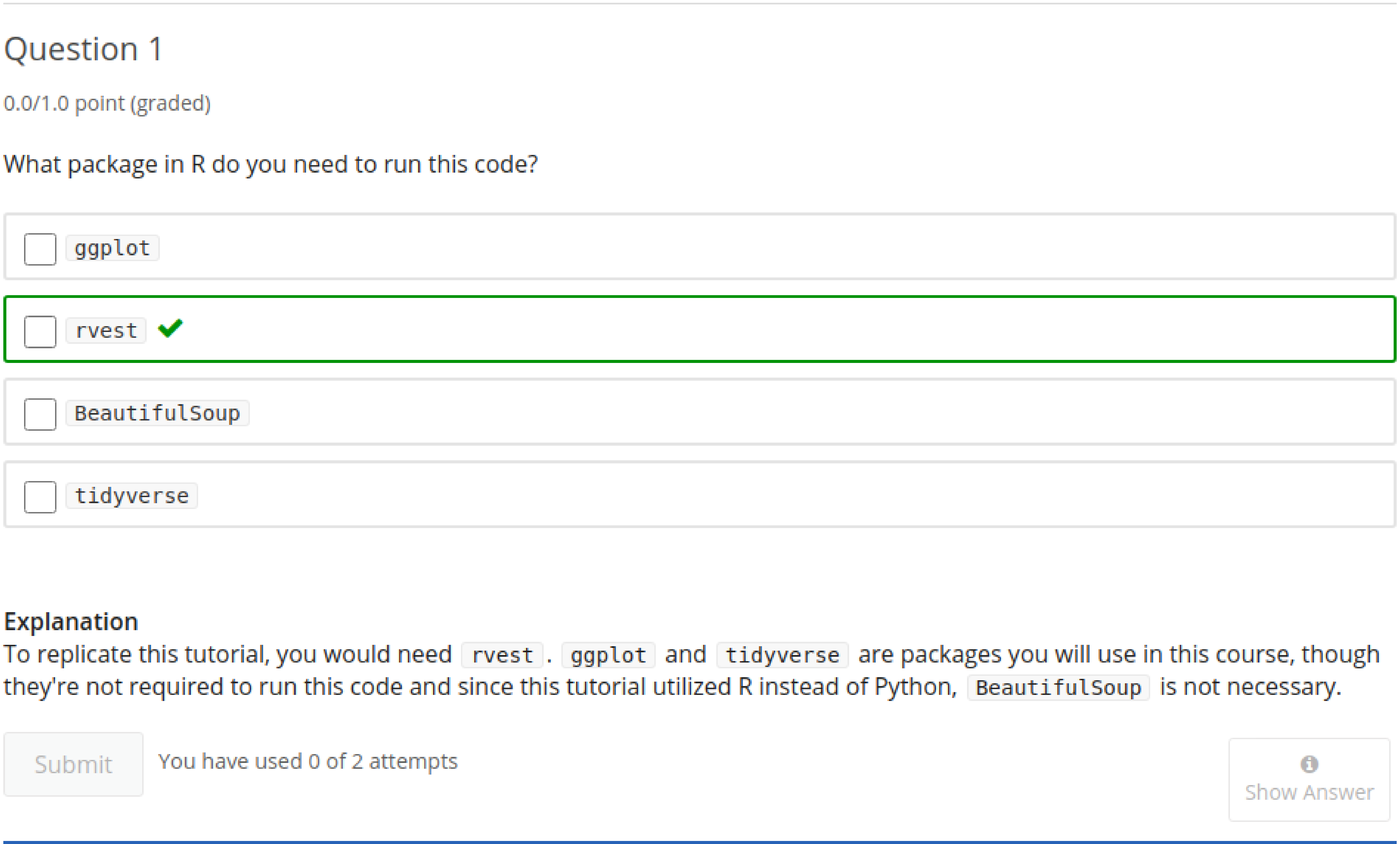
○ An API which helps users harvest data from major shopping sites, such as Amazon or AbeBooks

○ An online forum and tutorial site for discussing web scraping

○ A simple programming language meant for web scraping

Explanation

As Professor Duflo briefly demoes, BeautifulSoup is a simple Python package which allows users to find information on websites by utilizing a site's HTML or XML code.



Question 1

0.0/1.0 point (graded)

What package in R do you need to run this code?

□ `ggplot`

□ `rvest` ✓

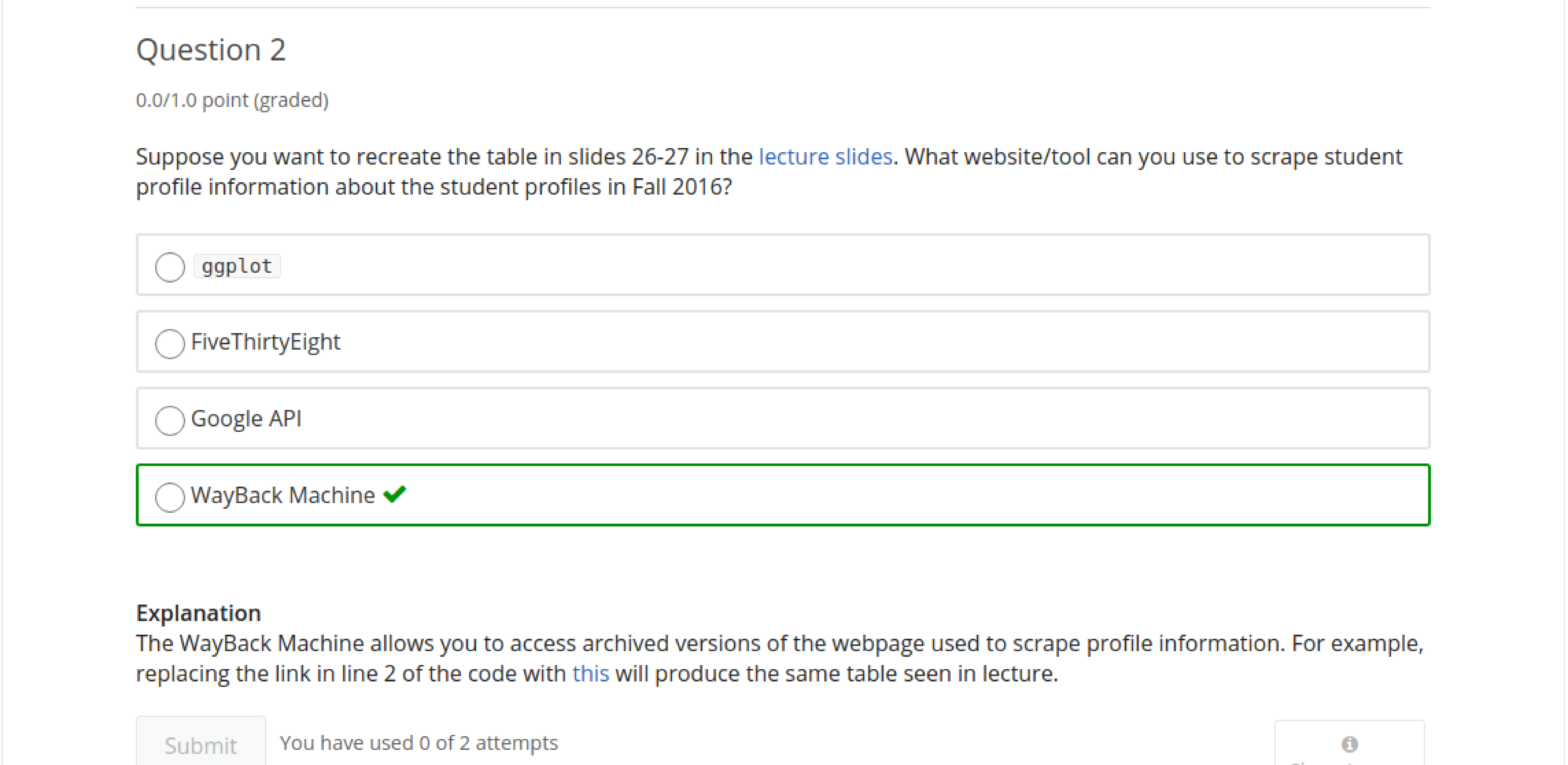
□ `BeautifulSoup`

□ `tidyverse`

Explanation

To replicate this tutorial, you would need `rvest`. `ggplot` and `tidyverse` are packages you will use in this course, though they're not required to run this code and since this tutorial utilized R instead of Python, `BeautifulSoup` is not necessary.

You have used 0 of 2 attempts



Question 2

0.0/1.0 point (graded)

Suppose you want to recreate the table in slides 26-27 in the lecture slides. What website/tool can you use to scrape student profile information about the student profiles in Fall 2016?

○ `ggplot`

○ FiveThirtyEight

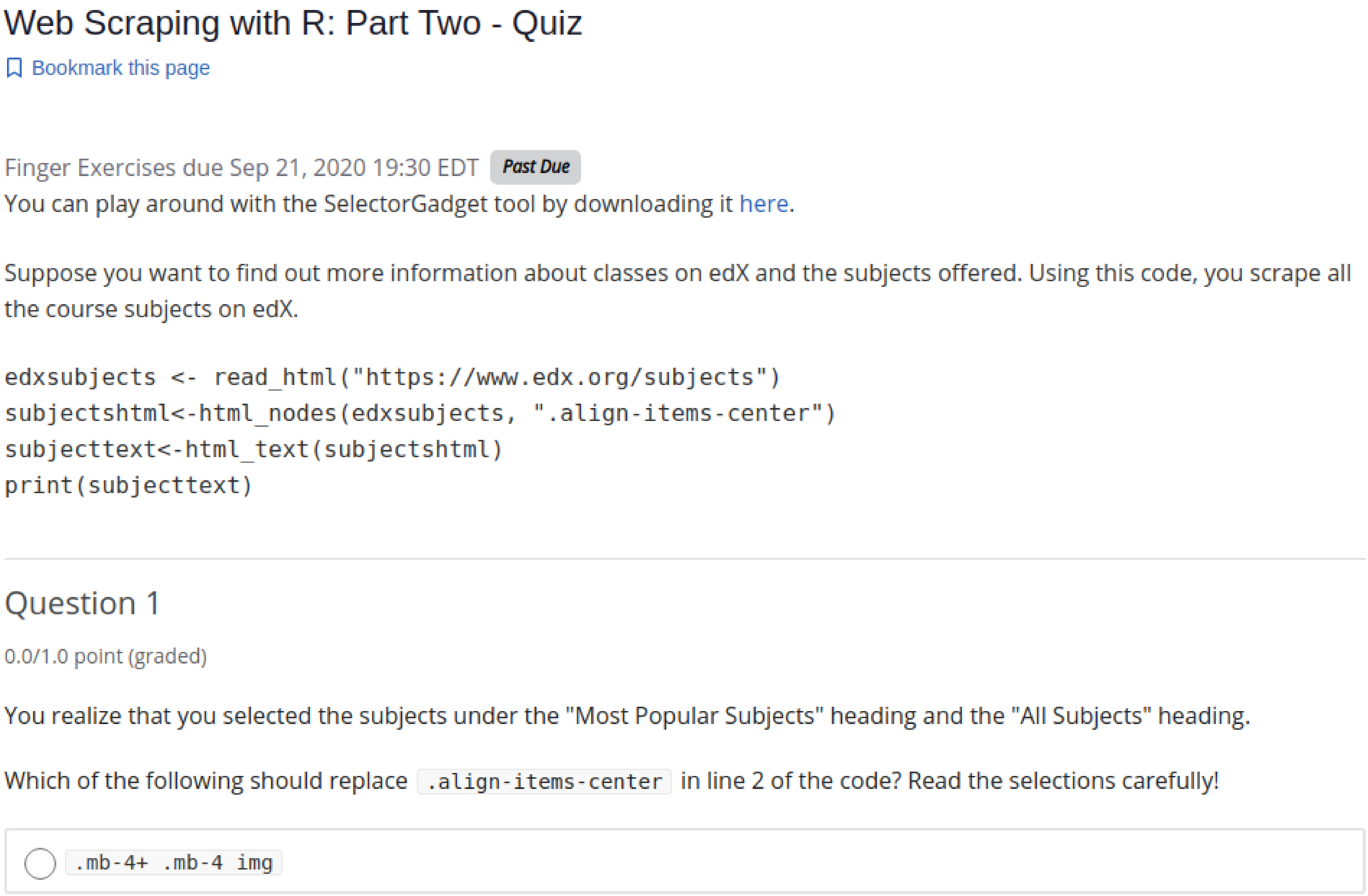
○ Google API

○ WayBack Machine ✓

Explanation

The WayBack Machine allows you to access archived versions of the webpage used to scrape profile information. For example, replacing the link in line 2 of the code with this will produce the same table seen in lecture.

You have used 0 of 2 attempts



Web Scraping with R: Part Two - Quiz

Bookmark this page

Finger Exercises due Sep 21, 2020 19:30 EDT **Past Due**

You can play around with the SelectorGadget tool by downloading it here.

Suppose you want to find out more information about classes on edX and the subjects offered. Using this code, you scrape all the course subjects on edX.

```
edxsubjects <- read_html("https://www.edx.org/subjects")
subjectshtml<-html_nodes(edxsubjects, ".align-items-center")
subjecttext<-html_text(subjectshtml)
print(subjecttext)
```

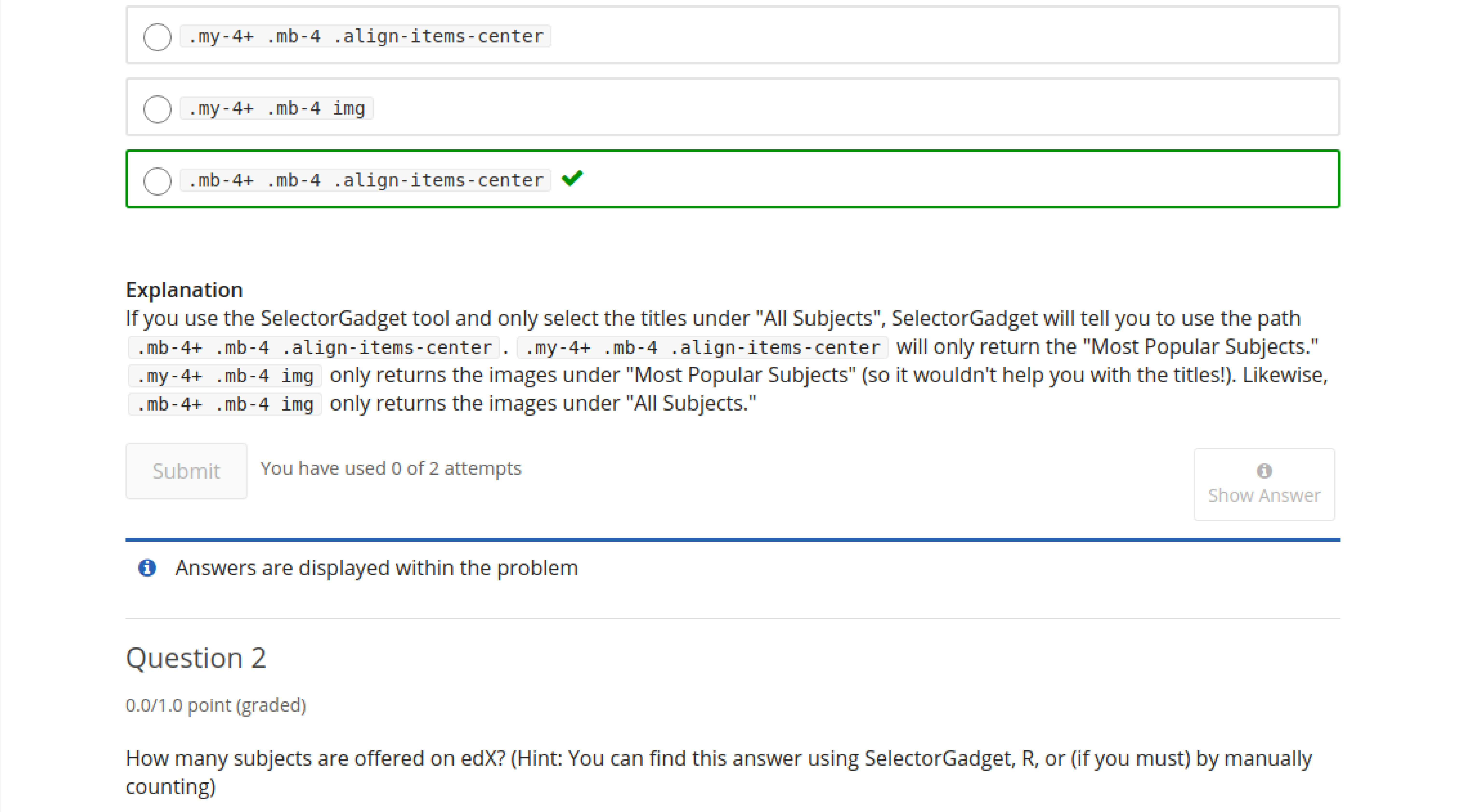
Question 1

0.0/1.0 point (graded)

You realize that you selected the subjects under the "Most Popular Subjects" heading and the "All Subjects" heading.

Which of the following should replace `.align-items-center` in line 2 of the code? Read the selections carefully!

○ `.mb-4+ .mb-4 img`



○ `.my-4+ .mb-4 .align-items-center`

○ `.my-4+ .mb-4 img`

○ `.mb-4+ .mb-4 .align-items-center` ✓

Explanation

If you use the SelectorGadget tool and only select the titles under "All Subjects", SelectorGadget will tell you to use the path `.mb-4+ .mb-4 .align-items-center`. `.my-4+ .mb-4 .align-items-center` will only return the "Most Popular Subjects." `.my-4+ .mb-4 img` only returns the images under "Most Popular Subjects" (so it wouldn't help you with the titles!). Likewise, `.mb-4+ .mb-4 img` only returns the images under "All Subjects."

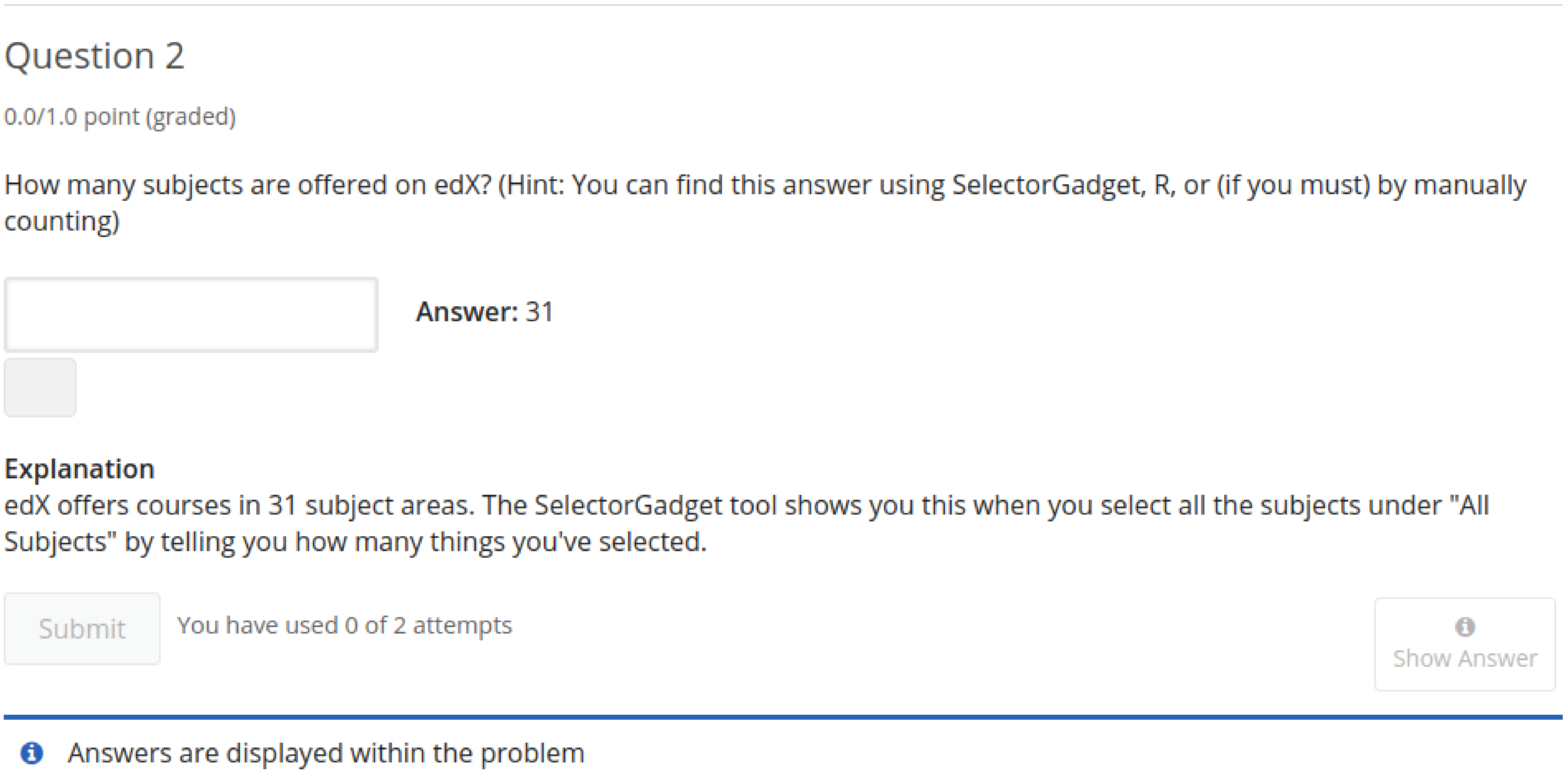
You have used 0 of 2 attempts

Answers are displayed within the problem

Question 2

0.0/1.0 point (graded)

How many subjects are offered on edX? (Hint: You can find this answer using SelectorGadget, R, or (if you must) by manually counting)



Question 2

0.0/1.0 point (graded)

How many subjects are offered on edX? (Hint: You can find this answer using SelectorGadget, R, or (if you must) by manually counting)

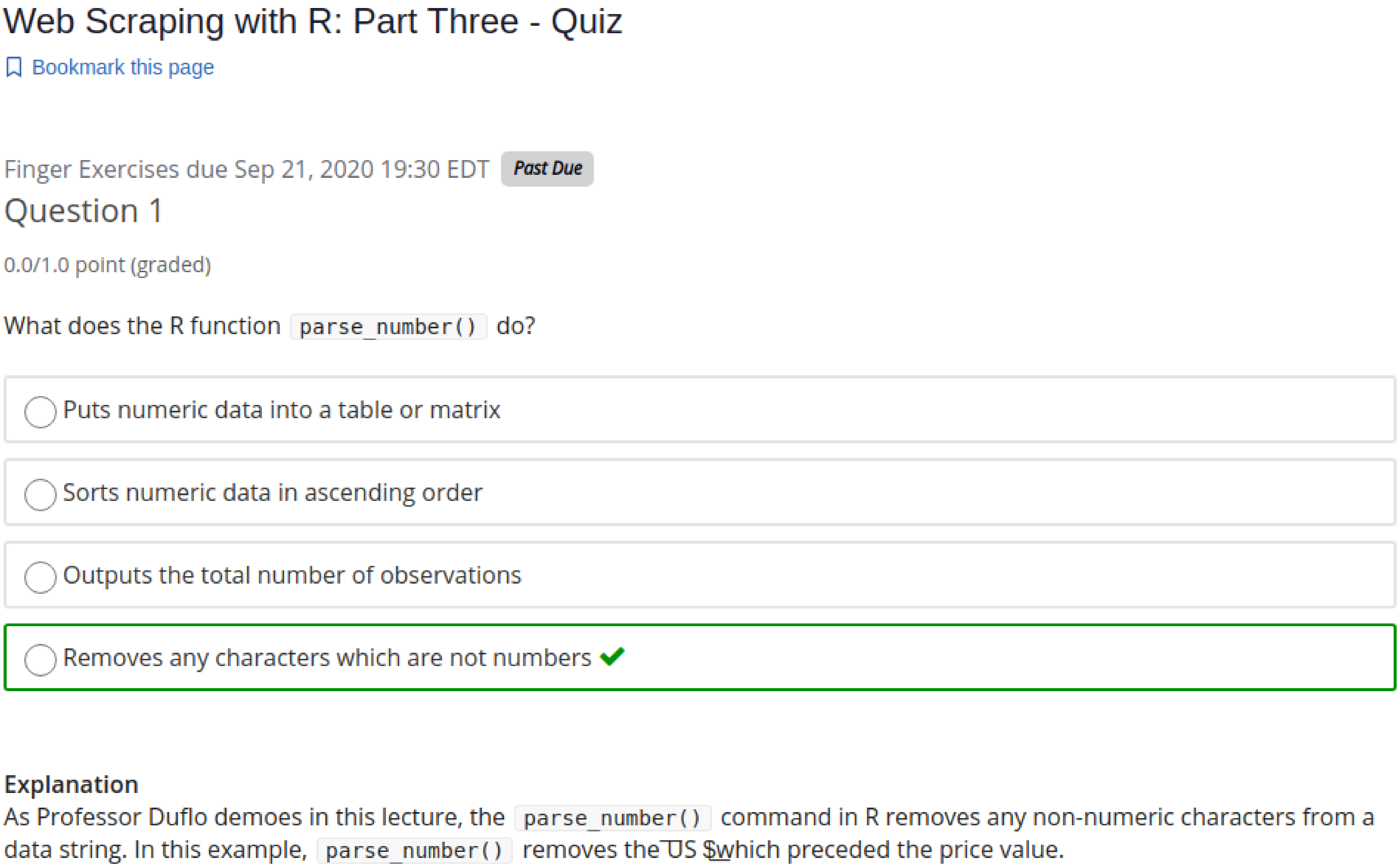
Answer: 31

Explanation

edX offers courses in 31 subject areas. The SelectorGadget tool shows you this when you select all the subjects under "All Subjects" by telling you how many things you've selected.

You have used 0 of 2 attempts

Answers are displayed within the problem



Web Scraping with R: Part Three - Quiz

Bookmark this page

Finger Exercises due Sep 21, 2020 19:30 EDT *Past Due*

Question 1

0.0/1.0 point (graded)

What does the R function `parse_number()` do?

○ Puts numeric data into a table or matrix

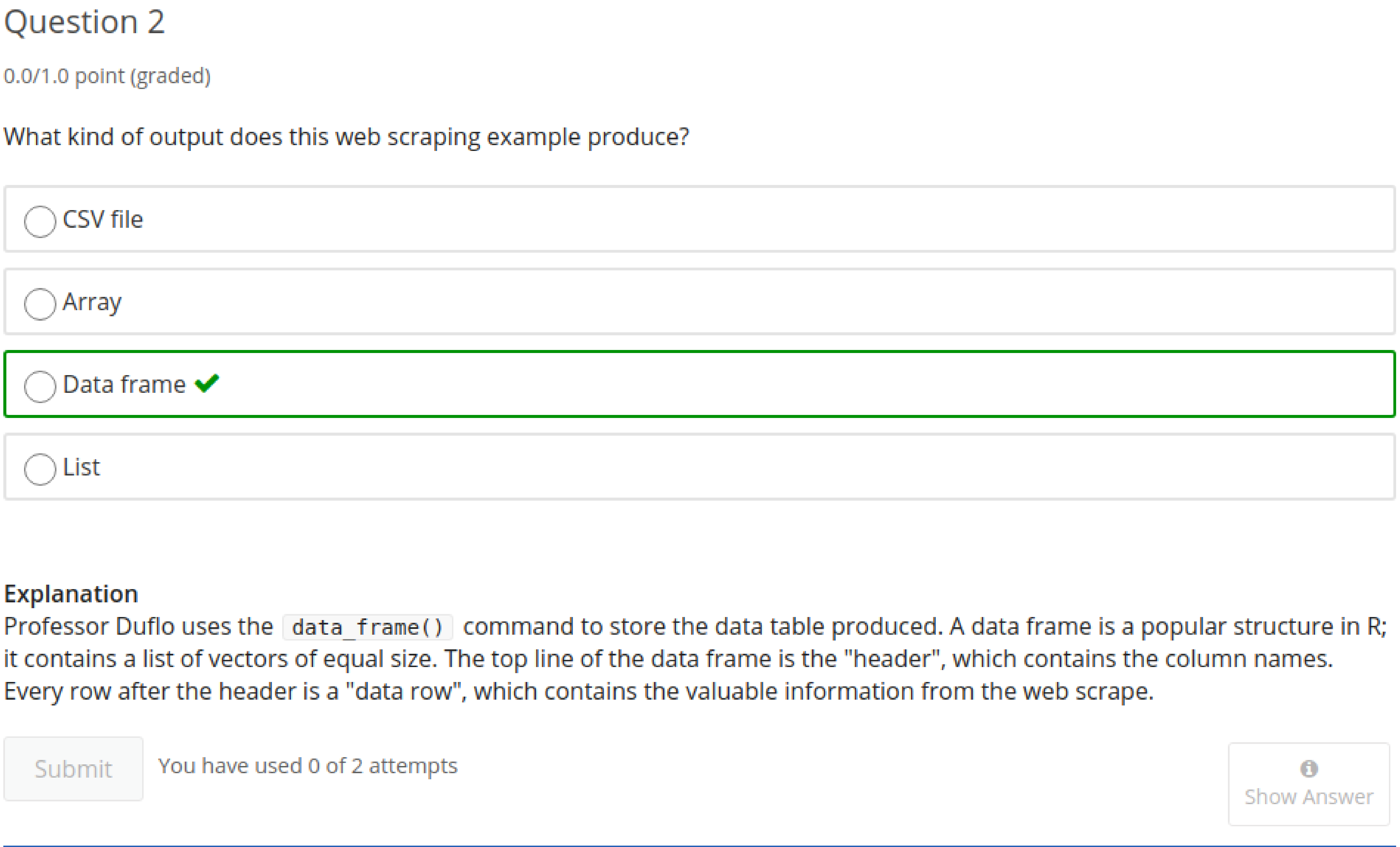
○ Sorts numeric data in ascending order

○ Outputs the total number of observations

○ Removes any characters which are not numbers ✓

Explanation

As Professor Duflo demoes in this lecture, the `parse_number()` command in R removes any non-numeric characters from a data string. In this example, `parse_number()` removes the US $which preceded the price value.



Question 2

0.0/1.0 point (graded)

What kind of output does this web scraping example produce?

○ CSV file

○ Array

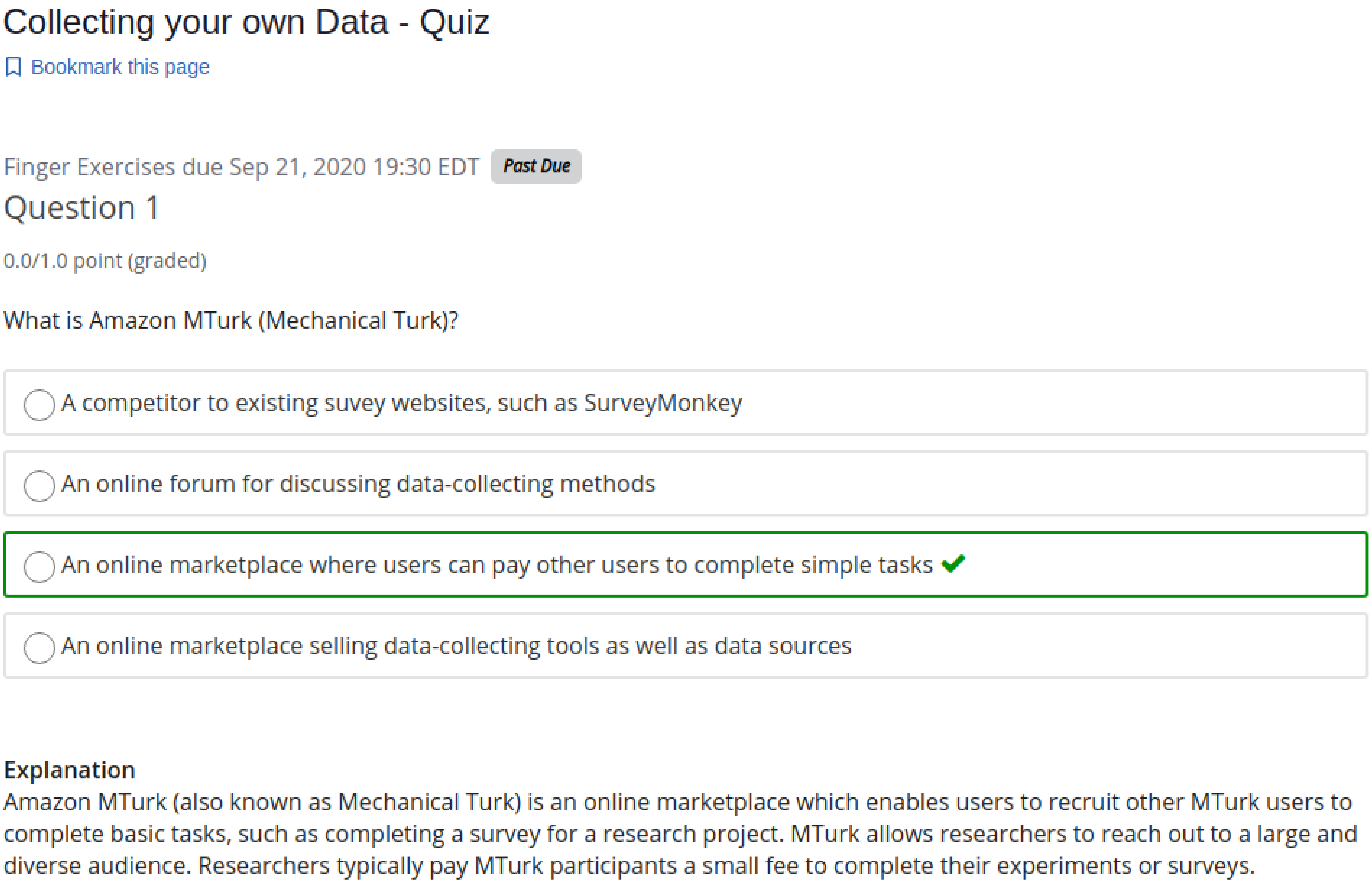
○ Data frame ✓

○ List

Explanation

Professor Duflo uses the `data_frame()` command to store the data table produced. A data frame is a popular structure in R; it contains a list of vectors of equal size. The top line of the data frame is the "header", which contains the column names. Every row after the header is a "data row", which contains the valuable information from the web scrape.

You have used 0 of 2 attempts



Collecting your own Data - Quiz

Bookmark this page

Finger Exercises due Sep 21, 2020 19:30 EDT **Past Due**

Question 1

0.0/1.0 point (graded)

What is Amazon MTurk (Mechanical Turk)?

○ A competitor to existing suvey websites, such as SurveyMonkey

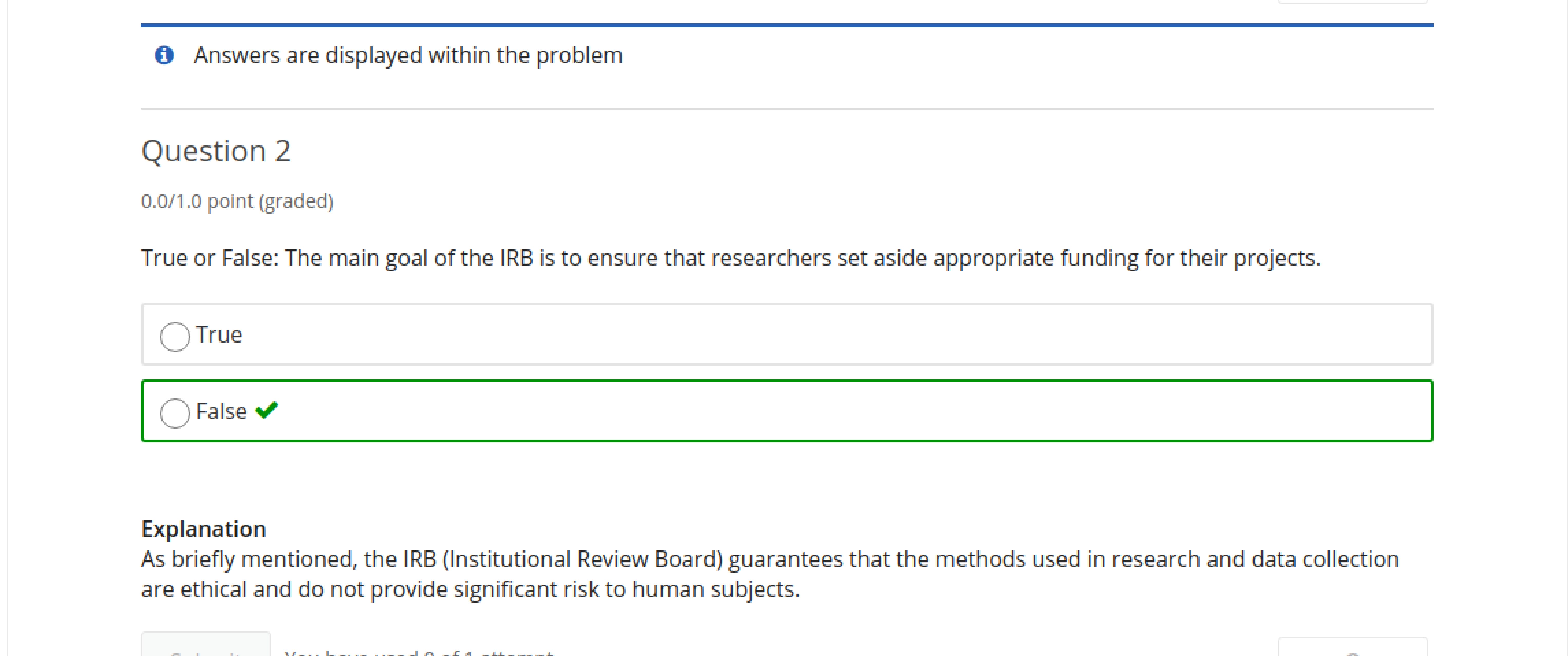
○ An online forum for discussing data-collecting methods

○ An online marketplace where users can pay other users to complete simple tasks ✓

○ An online marketplace selling data-collecting tools as well as data sources

Explanation

Amazon MTurk (also known as Mechanical Turk) is an online marketplace which enables users to recruit other MTurk users to complete basic tasks, such as completing a survey for a research project. MTurk allows researchers to reach out to a large and diverse audience. Researchers typically pay MTurk participants a small fee to complete their experiments or surveys.



Answers are displayed within the problem

Question 2

0.0/1.0 point (graded)

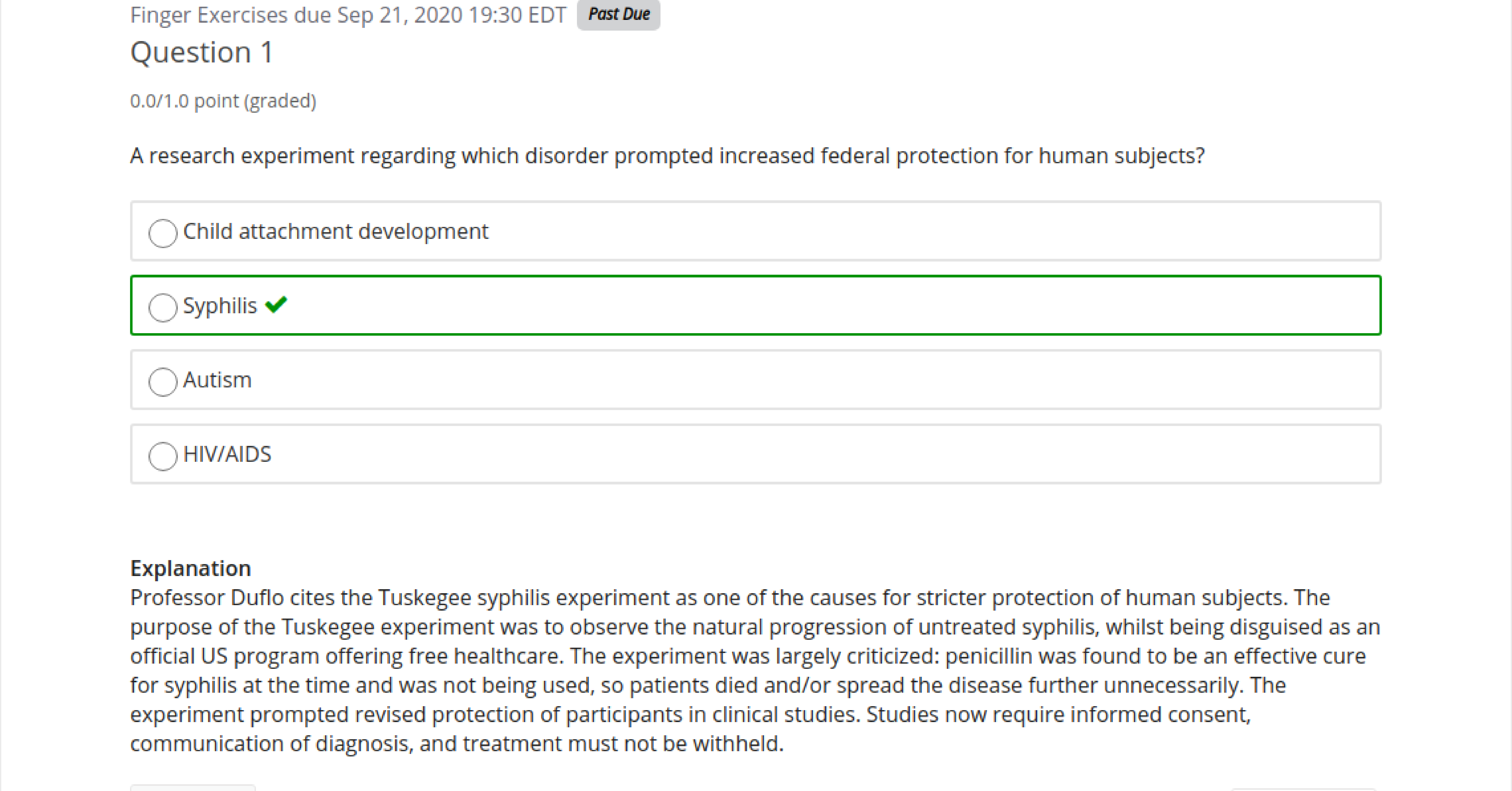
True or False: The main goal of the IRB is to ensure that researchers set aside appropriate funding for their projects.

○ True

○ False ✓

Explanation
As briefly mentioned, the IRB (Institutional Review Board) guarantees that the methods used in research and data collection are ethical and do not provide significant risk to human subjects.

You have used 0 of 1 attempt



Finger Exercises due Sep 21, 2020 19:30 EDT *Past Due*

Question 1

0.0/1.0 point (graded)

A research experiment regarding which disorder prompted increased federal protection for human subjects?

○ Child attachment development

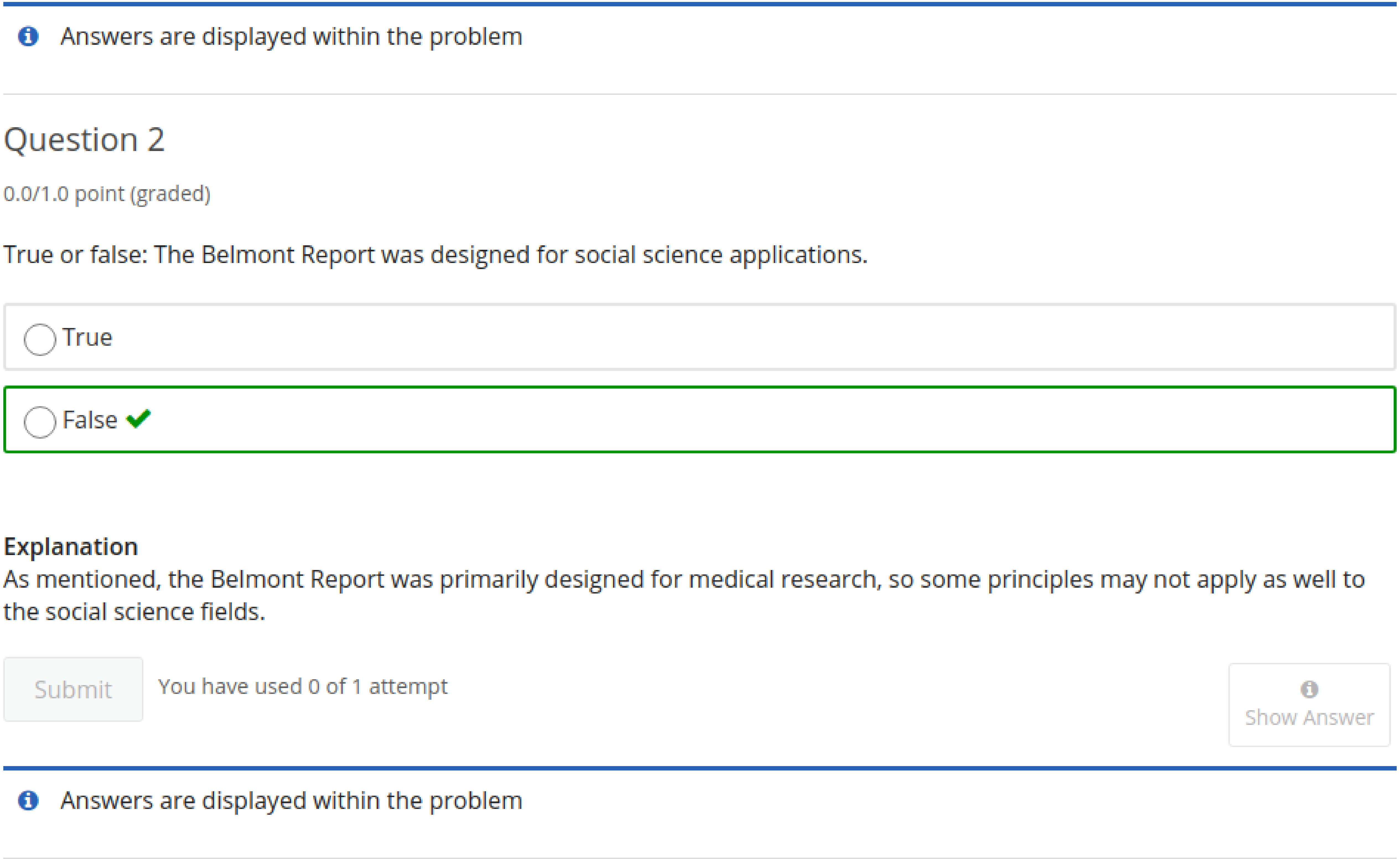
○ Syphilis ✓

○ Autism

○ HIV/AIDS

Explanation

Professor Duflo cites the Tuskegee syphilis experiment as one of the causes for stricter protection of human subjects. The purpose of the Tuskegee experiment was to observe the natural progression of untreated syphilis, whilst being disguised as an official US program offering free healthcare. The experiment was largely criticized: penicillin was found to be an effective cure for syphilis at the time and was not being used, so patients died and/or spread the disease further unnecessarily. The experiment prompted revised protection of participants in clinical studies. Studies now require informed consent, communication of diagnosis, and treatment must not be withheld.



Answers are displayed within the problem

Question 2

0.0/1.0 point (graded)

True or false: The Belmont Report was designed for social science applications.

○ True

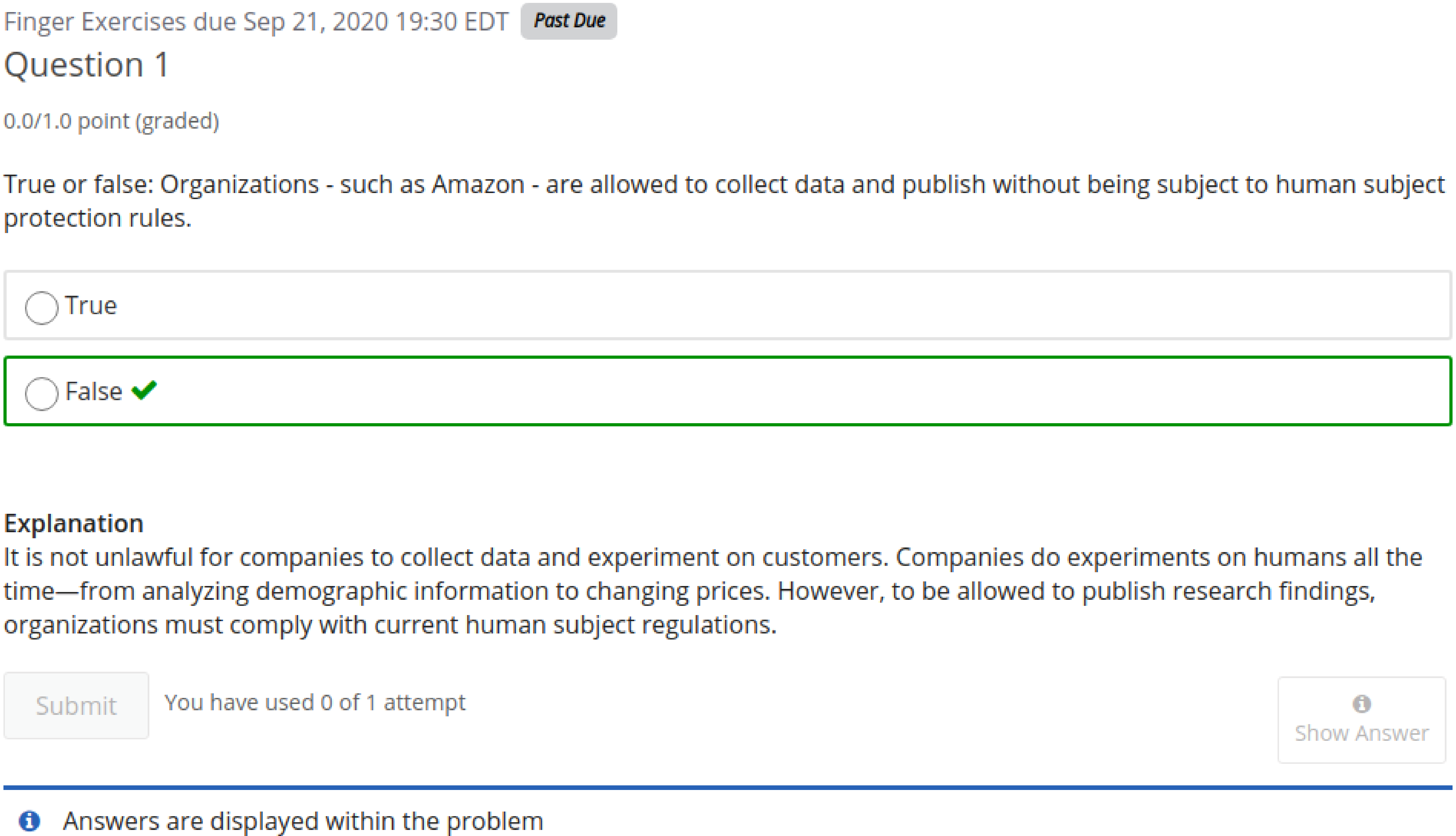
○ False ✓

Explanation

As mentioned, the Belmont Report was primarily designed for medical research, so some principles may not apply as well to the social science fields.

You have used 0 of 1 attempt

Answers are displayed within the problem



Finger Exercises due Sep 21, 2020 19:30 EDT Past Due

Question 1

0.0/1.0 point (graded)

True or false: Organizations - such as Amazon - are allowed to collect data and publish without being subject to human subject protection rules.

○ True

○ False ✓

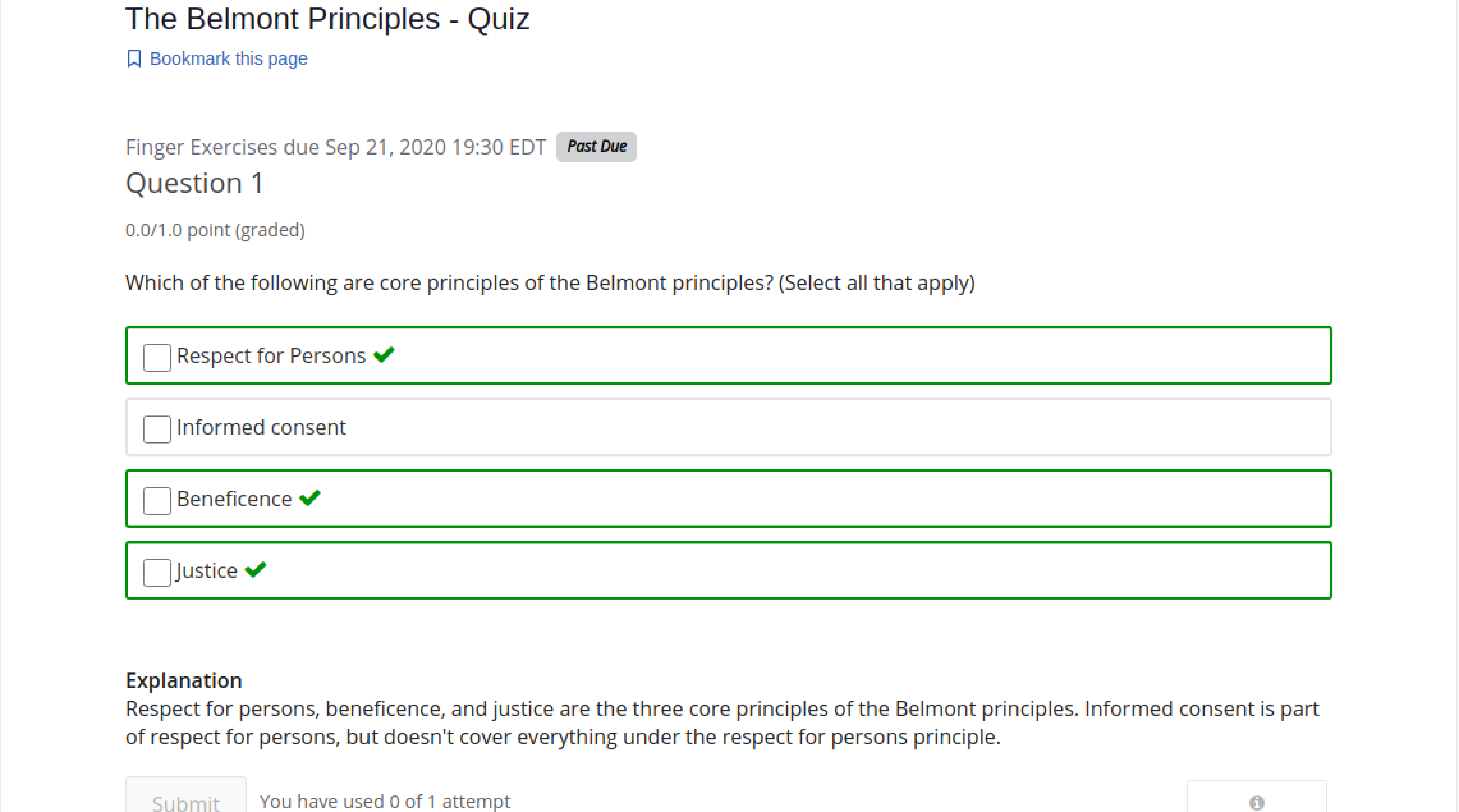
Explanation

It is not unlawful for companies to collect data and experiment on customers. Companies do experiments on humans all the time—from analyzing demographic information to changing prices. However, to be allowed to publish research findings, organizations must comply with current human subject regulations.

Submit You have used 0 of 1 attempt

Show Answer

Answers are displayed within the problem



The Belmont Principles - Quiz

Finger Exercises due Sep 21, 2020 19:30 EDT **Past Due**

Question 1

0.0/1.0 point (graded)

Which of the following are core principles of the Belmont principles? (Select all that apply)

☐ Respect for Persons ✓

☐ Informed consent

☐ Beneficence ✓

☐ Justice ✓

Explanation

Respect for persons, beneficence, and justice are the three core principles of the Belmont principles. Informed consent is part of respect for persons, but doesn't cover everything under the respect for persons principle.

You have used 0 of 1 attempt