Sem 1 Assessment

Rishi Valley 2018

Think & Write

Part 1: 9 Questions

Answer ALL Questions (not less that 50 words each)

1. How does data impact your life?
2. What is a trend?
3. Where does Google Trends data come from?
4. Why did Google Flu Trends eventually fail? What assumptions did they make about their data or their model that ultimately proved not to be true?
5. Why use charts and graphs rather than showing the raw data itself? (or) What are the benefits of visualizing data?
6. What makes a good/bad data visualization?
7. What are some reasons that you need to use a computer to manipulate data?
8. Name at least 6 types of charts you can use to visualize data.
9. Would you say that data analysis is a perfectly objective process? Why or why not?

Answers for these questions are in the slide decks we used in class.

Data Quiz

Part 2: 13 Multiple Choice Questions

Quiz 1

The digital divide is about how…

- **A.** …people's access to computing and digital technology increases over time through a process of dividing and growing quickly - it is often likened to the biological processes of cell growth
- **B.** …people's access to computing and the Internet differs based on socioeconomic or geographic characteristics.
- **C.** …people's access to computing technology is affected by the fact that newer devices that use new protocols makes it more difficult for them to communicate with older devices and technology
- **D.** …the amount of data on the Internet is growing so fast that the amount computing power and time we have to process it is lagging behind

Quiz 2

Which of the following is the most accurate statement about cleaning and filtering data?

- **A:** Using computing tools to filter and clean raw data makes it impossible to analyze or draw accurate conclusions
- **B:** Filtering and cleaning data is a fully automated process that should not require human input or intervention
- **C:** Filtering and cleaning data is a human process that does not require the use of computers
- **D:** Filtering and cleaning data is necessary to ensure that data is in a form that is better for computers to process

Quiz 3

Which of the following statements are true about pivot tables?

Select two answers.

- **A.** Pivot tables are used to quickly remove errors and inconsistencies from a dataset.
- **B.** Pivot tables are used to quickly perform aggregate computations and groupings on a set of raw data
- **C.** Pivot tables are used because they automatically detect and highlight potential trends or patterns in the underlying raw data
- **D.** Pivot tables are used to generate a summarized view of a large dataset which is helpful for gaining insight

Quiz 4

Biologists often attach tracking collars to wild animals. For each animal, the following geolocation data is collected at frequent intervals.

- The time,
- The date,
- The location of the animal

Which of the following questions about a particular animal could NOT be answered using only the data collected from the tracking collars?

- **A**. Approximately how many miles did the animal travel in one week?
- **B**. Does the animal travel in groups with other tracked animals?
- **C**. Do the movement patterns of the animal vary according to the weather?
- **D**. In what geographic locations does the animal typically travel?

Quiz 5

A bakery collects data on sales. Each sales record includes the date of the sale and some metadata about the items that were part of the sale.

The data includes: the names of the items sold, the types of items sold, the number of each item sold, and the price of each item sold.

Which of the following **CANNOT** be determined from the bakery's data set?

- **A.** The total income from sales the bakery received in the past month.
- **B.** Which customer most frequently purchases bread.
- **C.** The item bought in the highest quantity in the past week.
- **D.** Days when certain items sell the most.

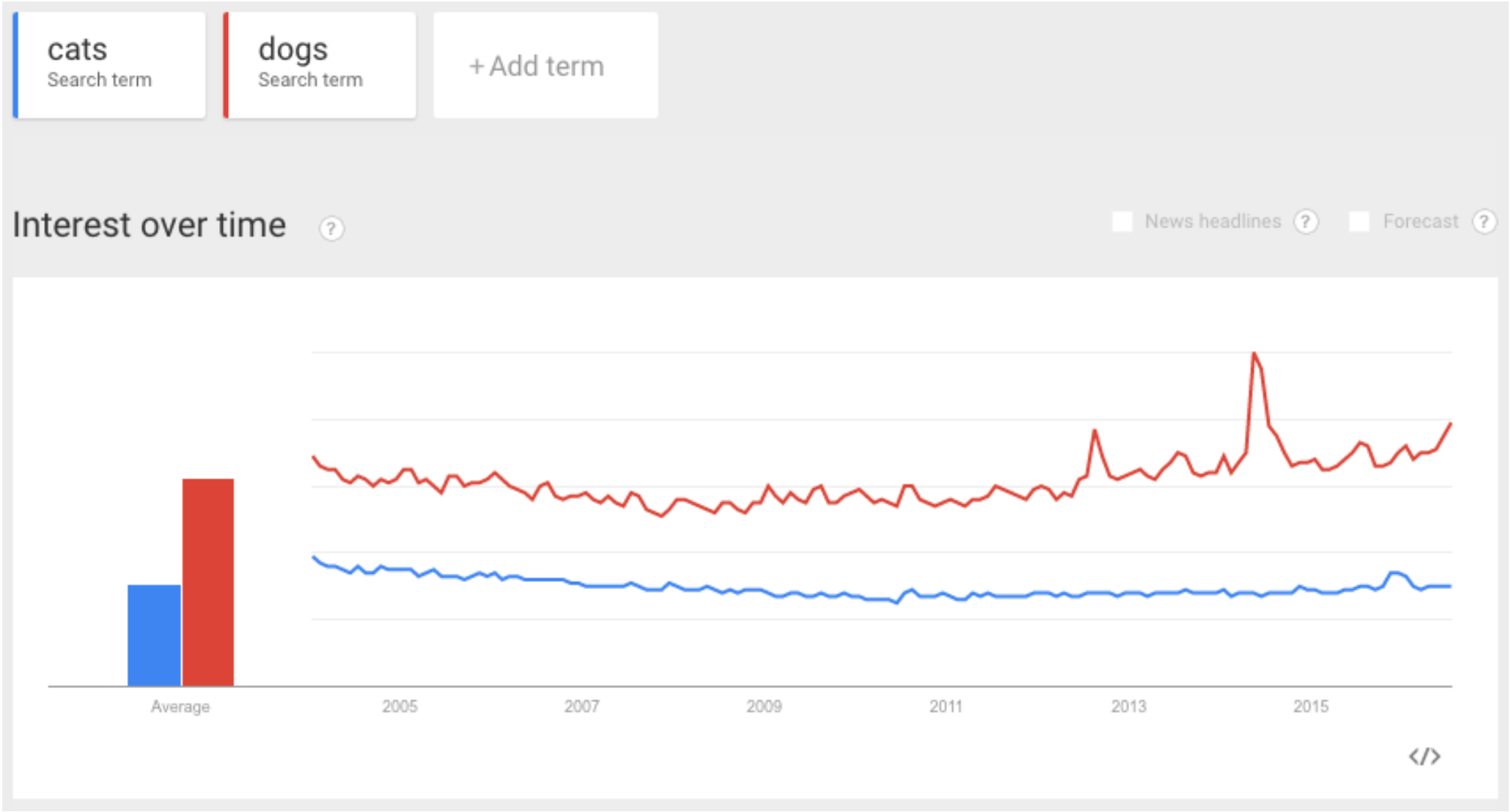
Quiz 6

- A certain social media Web site allows users to post messages and to comment on other messages that have been posted. When a user posts a message, the message itself is considered data. In addition to the data, the site stores the following meta data.
- The time the message was posted
- The name of the user who posted the message
- The names of users (and time) who comment on the message

For which of these goals is it more useful to analyze the data instead of the metadata?

- **A.** To determine the users who post messages most frequently
- **B.** To determine the time of day that the site is most active
- **C.** To determine the topics that many users are posting about
- **D.** To determine which posts from a user have received the greatest no of comments

Quiz 7


cats
Search term
dogs
Search term
+ Add term
Interest over time
News headlines
Forecast
Average
2005
2007
2009
2011
2013
2015

Quiz 7

Choose the most accurate description of what this data is actually showing based on what you know about how Google Trends works.

A. People like dogs more than cats

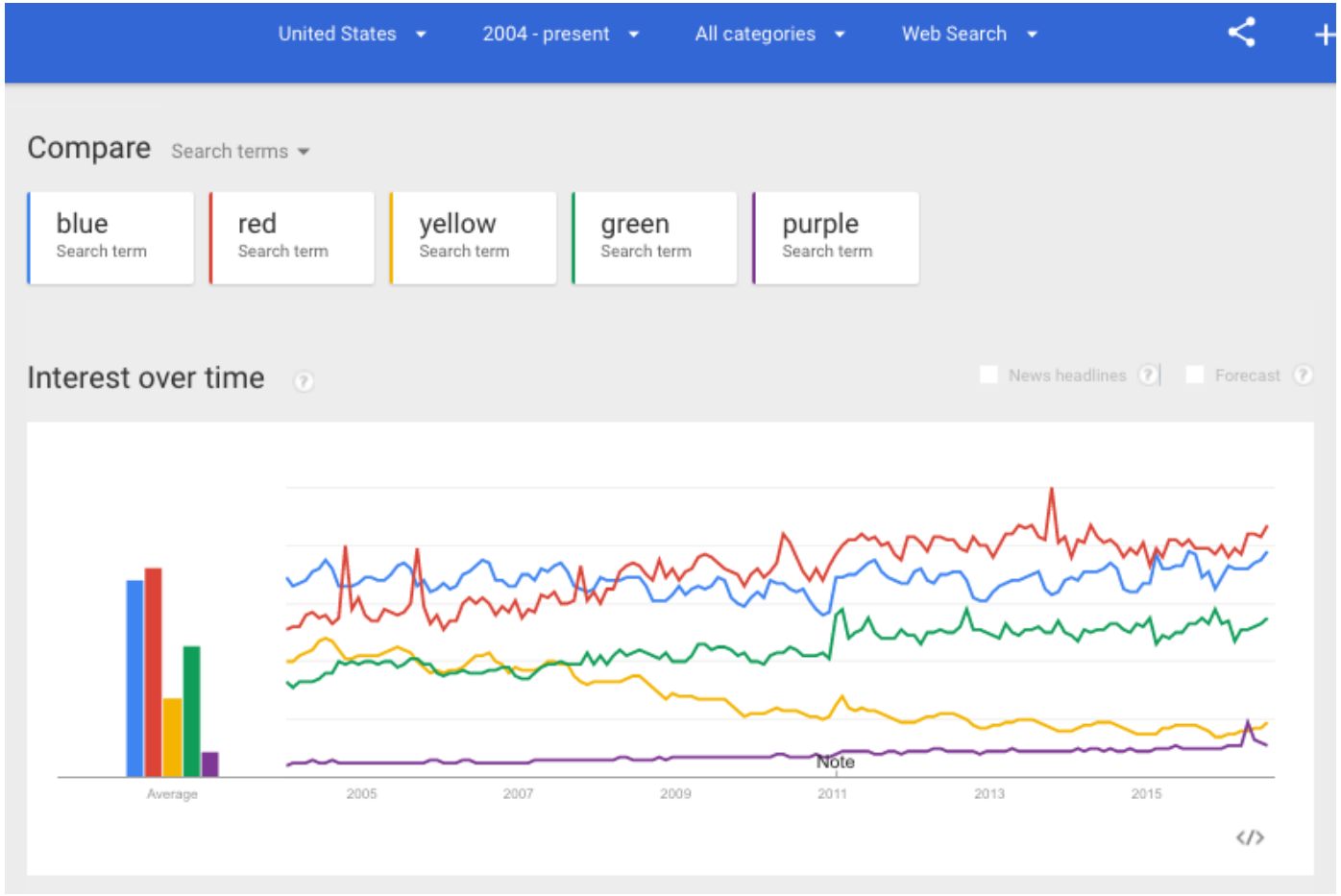
B. People search for "dogs" more frequently than "cats“

C. There was a sharp increase in the dog population between 2014 and 2015

D. The popularity of dogs as pets is slightly increasing over time, while the popularity of cats is relatively flat

Quiz 8

The Chart below from Google Trends shows the prevalence of some search terms in the United States between 2004 and the present. Which of the following is the most accurate statement of what this chart is showing.



Quiz 8

The Chart below from Google Trends shows the prevalence of some search terms in the United States between 2004 and the present. Which of the following is the most accurate statement of what this chart is showing.

- **A.** Since sometime around 2009, red has become the favorite color of more people
- **B.** Generally speaking, since 2009 more people use "red" in their search terms more than they use "blue", "yellow", "green", or "purple"
- **C.** The general decline in the search term "yellow" might be due to the decline of searches for yellow taxis, as car sharing services have become more popular
- **D.** Generally speaking, the volume of internet searches is increasing over time because the number of people using the internet is also increasing.

Quiz 9

Which of the following is the most accurate statement about using search trends as predictors of future events?

- **A.** Search trends are imperfect predictors of future events that fully represent society at large.
- **B.** Search trends are accurate and reliable predictors of future events that fully represent society at large.
- **C.** Search trends are imperfect predictors of future events that may not fully represent society at large.
- **D.** Search trends are accurate and reliable predictors of future events that may not fully represent society at large.

Quiz 10

The survey of high school seniors asked:

- What state do you live in?
- How likely are you to attend college in your home state? (on a scale of 1-5, 5 meaning "very likely")
- What do you plan to study?

Amara does an initial computation on the data to make a summary table. A small segment is shown below.

Home state	Likelihood of staying in state	Area of study	Num. of Responses
CA	2	English	5
Arkansas	3	Creative writing	2
MI	1.5	Applied Mathematics	10
Utah	2.5	Mathematics	7
adsfas	1	Adfa asdfa	1

Quiz 10

Amara is tasked with cleaning the data to prepare it for further analysis.

Which of the following would be the least appropriate modifications to make to the data to prepare it for further analysis?

- **A.** Translate all states into their two-letter state code
- **B.** Group similar areas of study into a single area of study. For example: grouping Applied Mathematics and Mathematics together into "Mathematics"
- C. Round up all non-integer values for "Likelihood of staying in state"
- **D.** Removing the entire row with home state "adsfas" and recomputing

Quiz 11

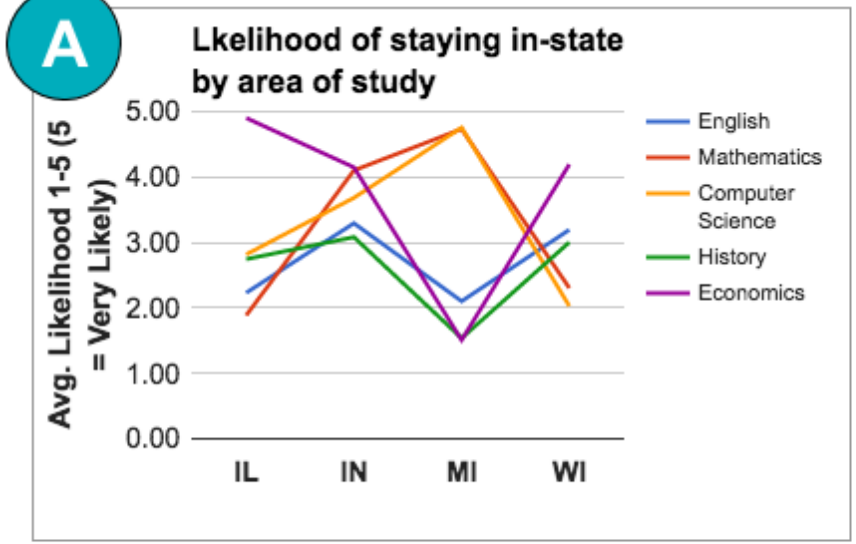
Amara plans to use the survey data to create a visualization and short write up about students' plans for college, but first she wants to learn more about how the survey was conducted. Of the following things she might learn about the survey, which are the most likely sources of bias in the results based how it was collected?

Choose two answers.

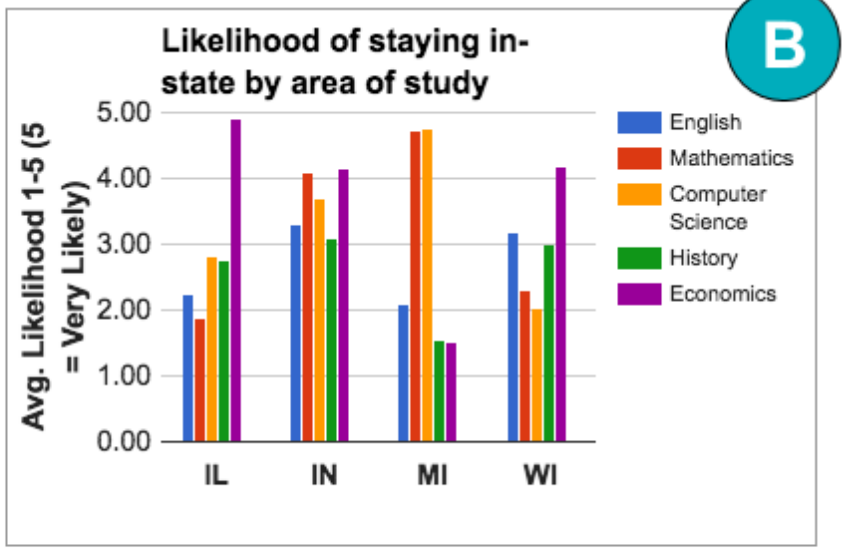
- **A.** She learns that the survey administrators only asked a representative sample of students, rather than every student in each state.
- B. She learns that responses were collected only by mobile app.
- C. She learns that the survey was only available to students who scored at the top 10% on the PSAT.
- **D.** She learns the survey was available to complete in both digital and paper form.

Quiz 13

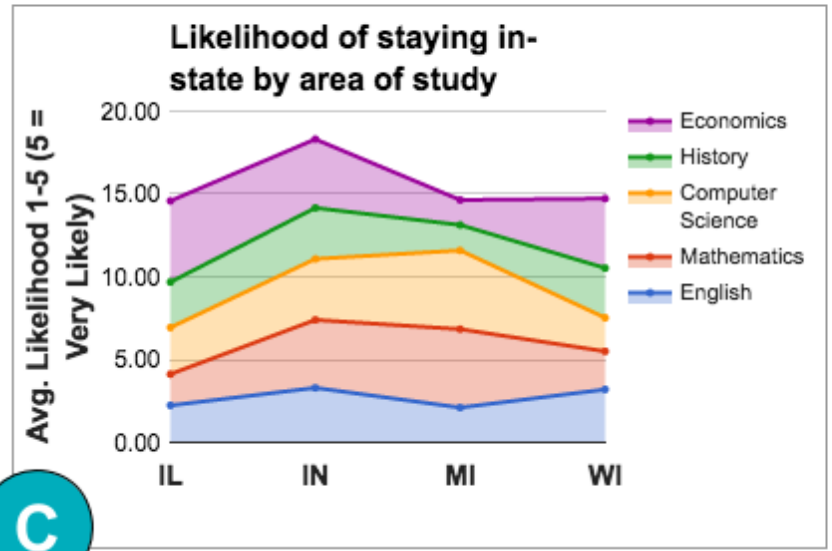
A


Lkelihood of staying in-state by area of study
Avg. Likelihood 1-5 (5 = Very Likely)
5.00
4.00
3.00
2.00
1.00
0.00
IL
IN
MI
WI
English
Mathematics
Computer Science
History
Economics

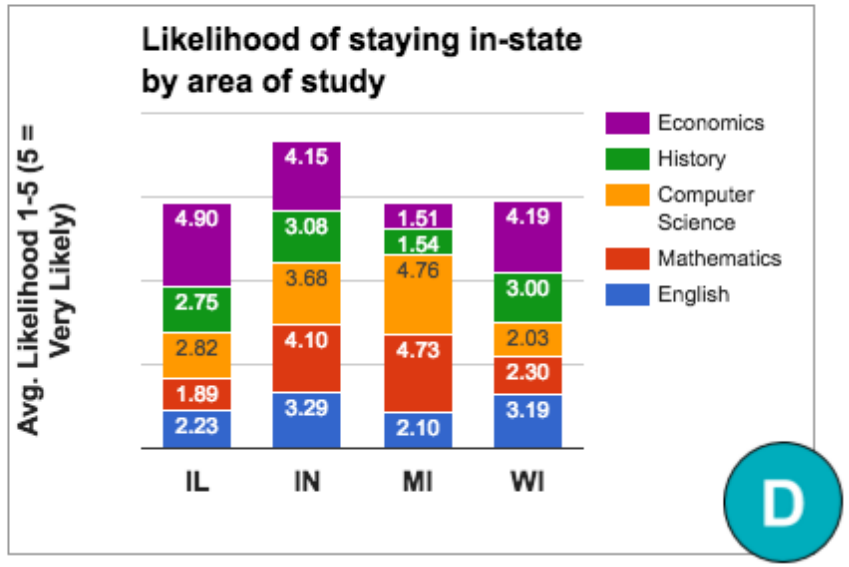
B


Likelihood of staying in-state by area of study
Avg. Likelihood 1-5 (5 = Very Likely)
5.00
4.00
3.00
2.00
1.00
0.00
IL
IN
MI
WI
English
Mathematics
Computer Science
History
Economics

C


Likelihood of staying in-state by area of study
Avg. Likelihood 1-5 (5 = Very Likely)
20.00
15.00
10.00
5.00
0.00
IL
IN
MI
WI
Economics
History
Computer Science
Mathematics
English

D


Likelihood of staying in-state by area of study
Avg. Likelihood 1-5 (5 = Very Likely)
4.90
2.75
2.82
1.89
2.23
4.15
3.08
3.68
4.10
3.29
1.51
1.54
4.76
4.73
2.10
4.19
3.00
2.03
2.30
3.19
IL
IN
MI
WI
Economics
History
Computer Science
Mathematics
English

Quiz 12

Amara decides to make a visualization of a portion of the responses showing only a few states and a few areas of study. She wants to make an effective visualization that shows for comparison: Students' average likelihood of attending college in-state broken down by which state they live in and what they plan to major in.

For example, in Illinois (IL) on average students who want to study economics are very likely to say they want to attend college in-state.

Amara makes four different visualizations shown below (marked A, B, C, D). According to good principles of visualization, and for what Amara wants to show, which one of these would be considered the best visual representation?

- **A.** Chart A (Line Chart)
- **B.** Chart B (Vertical Bar Chart)
- **C.** Chart C (Stacked Line Chart)
- **D.** Chart D (Stacked Vertical Bar Chart)

Quiz 12 Answer

B. Chart B (Vertical Bar Chart)

To see why B is correct it's easier to explain why the other responses are bad or at least not great...

- A is bad because connecting the points across states is meaningless. These are not trend lines, this is not showing change over time.
- C and D are both not great because stacked graphs attempt to show how parts contribute to a whole. So the problem is that the height of each stack doesn't mean anything - you can't add up averages this way. They are not as good visual representations because you have a key visual indicator - the height of the stacks - that doesn't actually convey any information. It is therefore potentially misleading because the viewer might make up their own meaning for the heights. A stack would be good if the data were about whole numbers of students rather than averages. We actually know nothing about the total number of students involved here.
- It's worth noting that D is accidentally not horrible, but only if you know what you're looking at, and know that each portion of the stack is basically independent of the stack itself -- it's arguably easier to compare the likelihoods across states.
- So B is the best choice here because there is nothing potentially misleading, and it also allows you to quickly see the answer to the intended question: for each state which types of students are most likely to stay in state? All of the other choices you have to do work to answer that question. In graph D for example, you just have to read the numbers, which defeats the purpose of having a chart - you could just look at the data.