

Dataforge — Activity Book

A Spark & Anvil activity book — puzzles, questions, and fun from Dataforge.

Write in the book, circle your answers, and check the Answer Key at the back.

1. Data Types & Collection

Circle the correct answer

1. What type of data describes categories or groups, such as favorite color or type of pet?

(A) Random data (B) Numerical data (C) Sequential data (D) Categorical data

2. What is a survey?

(A) A method of collecting data by asking people questions (B) A tool for measuring temperature (C) A type of graph used to display data (D) A mathematical formula for calculating averages

3. What type of data can be counted or measured with numbers, such as height or number of siblings?

(A) Categorical data (B) Descriptive data (C) Numerical data (D) Alphabetical data

4. What is the name for collecting data by watching and recording what happens without asking questions?

(A) Experiment (B) Observation (C) Interview (D) Survey

5. Why is it important to organize data into a table or chart after collecting it?

(A) You must always use a table or your data is invalid (B) Organizing data changes the values to make them more accurate (C) Tables make the data look prettier but don't help with understanding (D) Organizing data makes it easier to see patterns, compare values, and draw conclusions

6. How is collecting data through a survey different from collecting data through observation?

(A) Surveys can only be done online while observations are done in person (B) Surveys use numbers and observations use words (C) Observations are always more accurate than surveys (D) A survey asks people directly for information, while observation means watching and recording what happens without asking

Write the answer

7. Explain why 'favorite ice cream flavor' is categorical data and not numerical data.

Answer: _____

8. What does it mean for a data set to have variability?

Answer: _____

9. Why might the same survey question give different results when asked to two different groups of people?

Answer: _____

10. How would you explain the difference between a tally chart and a frequency table to a younger student?

Answer: _____

Match each question to its answer

1. Your class wants to find out the most popular lunch option in the cafeteria. Which data collection method would work best? →

2. You collected the following data about pets in your class: dog, cat, dog, fish, cat, dog, hamster, cat, dog, fish. How would you organize this into a frequency table? →

3. A scientist wants to know how many birds visit a park each hour. Should they use a survey or an observation? Why? →

4. Classify each of the following as categorical or numerical: (1) number of pages in a book, (2) color of a car, (3) temperature outside, (4) type of music. →

5. You surveyed your class and found that jersey numbers range from 1 to 99. Are jersey numbers categorical or numerical data? Explain your reasoning. →

(A) Numerical: pages in a book, temperature. Categorical: color of a car, type of music

(B) Observation, because you cannot ask birds questions but you can watch and count them

(C) Categorical, because jersey numbers are labels that identify players and it does not make sense to add or average them

(D) List each pet type with its count: Dog 4, Cat 3, Fish 2, Hamster 1

(E) A survey asking each student to choose their favorite lunch option from a list

2. Reading & Creating Graphs

Circle the correct answer

11. What type of graph uses rectangular bars to compare quantities?

(A) Line graph (B) Bar graph (C) Scatter plot (D) Pie chart

12. What is the horizontal line at the bottom of a graph called?

(A) The x-axis (B) The y-axis (C) The legend (D) The title

13. What type of graph shows data as parts of a whole circle?

(A) Line graph (B) Pie chart (C) Bar graph (D) Histogram

14. What is the purpose of a legend on a graph?

(A) To show the title of the graph (B) To display the scale of the axes (C) To list the source of the data (D) To explain what the colors or symbols on the graph represent

15. Why is a line graph better than a bar graph for showing temperature changes over a week?

(A) A line graph can only show temperature data (B) A line graph uses less ink than a bar graph (C) A line graph shows how data changes continuously over time (D) A line graph always makes numbers look larger

16. What does the scale on a graph's y-axis tell you?

(A) The date when the data was collected (B) The range and intervals used to measure the data values (C) The name of the person who made the graph (D) The number of categories in the data

Write the answer

17. Why is it important that a bar graph's y-axis starts at zero?

Answer: _____

18. How is a pictograph different from a bar graph?

Answer: _____

19. What does a steep upward line on a line graph indicate?

Answer: _____

20. How would you explain to a friend why every graph needs a title?

Answer: _____

Match each question to its answer

1. A bar graph shows book sales: Fiction = 40, Nonfiction = 25, Comics = 35, Poetry = 10. What is the total number of books sold? →

2. A pictograph shows apples picked: Monday has 4 apple symbols, Tuesday has 6, Wednesday has 3. Each symbol represents 5 apples. How many apples were picked on Tuesday? →

3. You surveyed classmates about favorite seasons: Spring = 7, Summer = 12, Fall = 5, Winter = 6. Which type of graph would best display this data? →

4. A line graph tracks plant growth over 5 weeks: Week 1 = 2 cm, Week 2 = 4 cm, Week 3

= 7 cm, Week 4 = 9 cm, Week 5 = 10 cm. During which week did the plant grow the most? →

5. A pie chart shows how students spend after-school time: Homework 30%, Sports 25%, Screen Time 20%, Reading 15%, Other 10%. If 200 students were surveyed, how many chose Sports? →

- (A) 110 books
 - (B) 30 apples
 - (C) 50 students
 - (D) A bar graph because it compares quantities across categories
 - (E) Between Week 2 and Week 3, when it grew 3 cm
-

3. Measures of Center

Circle the correct answer

21. What is the mean of a data set?

- (A) The middle value when data is ordered (B) The value that appears most often
(C) The difference between the highest and lowest values (D) The sum of all values divided by the number of values

22. What is the median of a data set?

- (A) The middle value when the data is arranged in order (B) The average of all the values
(C) The most common value in the data (D) The largest value minus the smallest value

23. What is the mode of a data set?

- (A) The value that is exactly in between the highest and lowest (B) The value that appears most frequently
(C) The middle value when data is ordered (D) The sum of all values divided by the count

24. What is the range of a data set?

- (A) The middle value in the ordered data (B) The difference between the highest and lowest values
(C) The number of values in the data set (D) The average of all the numbers

25. Why might the mean not be a good way to describe a data set that contains an outlier?

- (A) Because the mean is never accurate for any data set (B) An outlier pulls the mean away from where most of the data is clustered
(C) Because outliers make the mean equal to zero (D) Because the mean cannot be calculated when there is an outlier

- 26.** In what situation would the median be a better measure of center than the mean?
(A) When the data set has an odd number of values (B) When the data has extreme values that would skew the average (C) When all the numbers in the data set are the same (D) When you want to find the most common value

Write the answer

- 27.** How would you explain the difference between mean and median to a younger student?

Answer: _____

- 28.** Can a data set have no mode? Explain why or why not.

Answer: _____

- 29.** Why are mean, median, and mode all called 'measures of center'?

Answer: _____

- 30.** What does it mean when a data set is 'skewed'?

Answer: _____

Match each question to its answer

1. Find the mean of this data set: 12, 15, 18, 21, 24. →
2. Find the median of this data set: 7, 3, 9, 1, 5, 8, 2. →
3. Find the mode of this data set: 4, 7, 2, 7, 5, 3, 7, 5, 9. →
4. Find the median of this data set with an even number of values: 10, 14, 18, 22. →
5. A student scored 85, 90, 78, 92, and 95 on five tests. What score does the student need on the 6th test to have a mean of 90? →

- (A) 100
 - (B) 18
 - (C) 5
 - (D) 7
 - (E) 16
-

4. Probability

Circle the correct answer

- 31.** What is probability?

- (A) The average of a set of numbers (B) A guarantee that something will happen
(C) The number of times something has happened in the past (D) A measure of how likely an event is to happen

32. What is the probability of an event that is impossible?

- (A) 0 (B) 1 (C) -1 (D) 0.5

33. What is the difference between theoretical probability and experimental probability?

- (A) Theoretical probability is calculated using math; experimental probability is found by actually doing an experiment and recording results (B) Experimental probability uses formulas while theoretical probability uses guessing (C) There is no difference; they always give the same answer (D) Theoretical probability is always larger than experimental probability

34. What does 'equally likely outcomes' mean?

- (A) All outcomes will happen the same number of times (B) Each outcome has the same chance of occurring (C) There are always exactly two possible outcomes (D) The outcomes are listed in order from least to most likely

35. Why does flipping a fair coin many times tend to give results closer to 50% heads and 50% tails?

- (A) The coin remembers previous flips and tries to even out, and this has been consistently demonstrated in practice (B) After many flips, the coin becomes more balanced, which is why this approach is recommended by experts (C) You are more likely to flip fairly after practice, making this the most widely accepted explanation (D) The Law of Large Numbers says that as you repeat an experiment more times, the experimental probability gets closer to the theoretical probability

36. A bag contains 3 red marbles and 7 blue marbles. Why is drawing a blue marble more likely than drawing a red marble?

- (A) Red marbles are harder to grab (B) There are more blue marbles than red marbles, so blue has a greater fraction of the total (C) The probability of any color is always 50% (D) Blue marbles are heavier, so they settle to the bottom

Write the answer

37. If you roll a standard die and get a 6 three times in a row, what is the probability of getting a 6 on the next roll?

Answer: _____

38. What does it mean when we say two events are 'complementary'?

Answer: _____

39. Explain why a weather forecast that says '70% chance of rain' does NOT mean it will rain for 70% of the day.

Answer: _____

40. How is probability related to fractions?

Answer: _____

Match each question to its answer

1. A spinner has 4 equal sections: red, blue, green, and yellow. What is the probability of landing on blue? →
2. A standard deck of cards has 52 cards. There are 4 aces. What is the probability of randomly drawing an ace? →
3. You flip a coin 50 times and get heads 28 times. What is the experimental probability of getting heads? →
4. A bag has 5 red, 3 blue, and 2 green marbles. What is the probability of NOT drawing a red marble? →
5. You roll two standard dice. What is the probability that the sum equals 7? →

- (A) $\frac{4}{52}$ or $\frac{1}{13}$
 - (B) $\frac{6}{36}$ or $\frac{1}{6}$
 - (C) $\frac{28}{50}$ or 56%
 - (D) $\frac{5}{10}$ or $\frac{1}{2}$
 - (E) $\frac{1}{4}$ or 25%
-

5. Data Distributions

Circle the correct answer

41. What does a dot plot use to represent each data value?
(A) A connected line between points (B) A bar whose height shows the value (C) A dot (or point) placed above a number line (D) A slice of a circle
42. What is the name for the middle value when all data values are listed in order?
(A) Range (B) Median (C) Mean (D) Mode
43. In a histogram, what does the height of each bar represent?
(A) The frequency (count) of data values in that interval (B) The exact data value
(C) The range of the entire data set (D) The average of values in that interval
44. What are the five values shown in a box plot's five-number summary?
(A) Minimum, Q1, median, Q3, and maximum (B) Q1, Q2, Q3, Q4, and Q5 (C)
Minimum, mean, median, mode, and maximum (D) Mean, median, mode, range, and count

45. A dot plot of test scores shows most dots clustered between 80 and 90, with a few dots near 50. What does this tell you about the data?

- (A) The average score is exactly 70 (B) The scores are evenly spread from 50 to 90
(C) Every student scored above 50 (D) Most students scored between 80 and 90, but a few scored much lower

46. A histogram is taller on the left side and has a long tail stretching to the right. What is the shape of this distribution?

- (A) Skewed right (positively skewed) (B) Uniform (C) Symmetric (D) Skewed left (negatively skewed)

Write the answer

47. Why is the interquartile range (IQR) often preferred over the full range as a measure of spread?

Answer: _____

48. In a symmetric distribution, how do the mean and median compare?

Answer: _____

49. A distribution has two distinct peaks. What term describes this shape?

Answer: _____

50. What does a wider box in a box plot tell you compared to a narrower box?

Answer: _____

Match each question to its answer

1. A data set contains these values: 12, 15, 14, 13, 16, 14, 15, 13, 14, 15. Create a dot plot description and identify the mode. →

2. The five-number summary of a data set is: Min = 20, Q1 = 35, Median = 50, Q3 = 65, Max = 80. What is the IQR? →

3. You have data on the number of books 20 students read last month: values range from 0 to 12. Which display would best show the shape of this distribution? →

4. A data set has a mean of 72 and a median of 68. In which direction is the distribution likely skewed? →

5. Construct a histogram with intervals 0-9, 10-19, 20-29, 30-39 for this data: 5, 12, 23, 7, 31, 15, 28, 3, 19, 22. How many values fall in the 20-29 interval? →

(A) 30

(B) Skewed right, because the mean is pulled higher than the median by high values

(C) 14 and 15 each appear three times, making the data bimodal at 14 and 15

(D) A dot plot, because the data set is small and values are whole numbers

(E) 3

6. Statistical Questions

Circle the correct answer

51. What makes a question a 'statistical question'?

(A) It is a question that uses numbers (B) It is a question that anticipates variability in the data collected to answer it (C) It is a question asked by a statistician (D) It is a question that has exactly one correct answer

52. What is a population in statistics?

(A) The average value in a data set (B) The entire group of individuals or items that you want to study (C) The number of people in a city (D) A small group selected for a survey

53. What is a sample in statistics?

(A) A single data point from the study (B) A free trial of a product (C) The entire group being studied (D) A subset of the population selected to represent the whole group

54. What does 'random sampling' mean?

(A) You choose people who volunteer for the study (B) You pick people who happen to be nearby (C) Every member of the population has an equal chance of being selected (D) You select only people who are easy to contact

55. Which of these is a statistical question?

(A) What is the school principal's name? (B) What is 7 times 8? (C) How many hours per week do students at our school spend on homework? (D) How many days are in February this year?

56. A student asks: 'How old is my math teacher?' Is this a statistical question? Why or why not?

(A) Yes, because you could ask multiple people and get different guesses (B) No, because it has a single, fixed answer with no variability (C) Yes, because age is a number and statistics involves numbers (D) No, because age is not an important topic in statistics

Write the answer

57. Why is it important that a sample be representative of the population?

Answer: _____

58. What is the difference between a survey and an experiment?

Answer: _____

59. What is sampling bias?

Answer: _____

60. Why is variability considered the heart of statistics?

Answer: _____

Match each question to its answer

1. A school wants to know students' favorite lunch options. They survey every 5th student entering the cafeteria. What type of sampling method is this? →
2. Rewrite the non-statistical question 'What is the temperature outside right now?' to make it a statistical question. →
3. You want to find out how much time middle schoolers in your state spend playing video games per week. Your population is too large to survey everyone. Describe how you would select a random sample of 100 students. →
4. A survey asks: 'Don't you agree that our school lunches are terrible?' What type of bias does this question introduce? →
5. A company surveys its employees about job satisfaction by only asking managers. Identify the problem and suggest a fix. →

(A) Response bias (leading question) — the wording pushes respondents toward a negative answer

(B) Systematic sampling

(C) Get a list of all middle schoolers in the state, assign each a number, then use a random number generator to select 100

(D) What are the daily high temperatures in our city over the past month?

(E) The sample is biased because managers may have different satisfaction levels than other employees; survey a random sample from all job levels

7. Data Analysis

Circle the correct answer

61. What does MAD stand for in statistics?

(A) Mean Absolute Deviation (B) Mean Adjusted Data (C) Maximum Absolute Distance (D) Median Average Difference

62. What type of graph uses two back-to-back stem-and-leaf displays to compare two data sets?

(A) A back-to-back stem-and-leaf plot (B) A scatter plot (C) A double bar graph (D) A side-by-side histogram

63. What does it mean when we say two data sets have 'overlapping distributions'?

(A) The ranges of the two data sets share common values (B) The two data sets are identical (C) The means of the two data sets are equal (D) The data was collected at the same time

64. In a scatter plot, what does it mean when the points form a pattern going up from left to right?

(A) There is a negative association between the variables (B) The variables are unrelated (C) One variable causes the other to increase (D) There is a positive association between the two variables

65. A data set of test scores has a mean of 78 and a MAD of 6. What does the MAD tell you about this data?

(A) The range of the data is 6 points (B) Exactly 6 students scored above the mean (C) The lowest score is 6 and the highest is 78 (D) On average, individual scores are about 6 points away from the mean of 78

66. Why might you choose to display data using side-by-side box plots instead of two separate histograms?

(A) Histograms cannot show two data sets at once (B) Box plots are always more accurate than histograms (C) Box plots show individual data values while histograms do not (D) Box plots make it easier to directly compare center, spread, and shape between groups on the same scale

Write the answer

67. What does it mean when a scatter plot shows 'no association' between two variables?

Answer: _____

68. How does the shape of a distribution affect which measure of center is most appropriate?

Answer: _____

69. When comparing two groups, what does it mean to express the difference between their centers in terms of MAD?

Answer: _____

70. What is an 'informal inference' in statistics?

Answer: _____

Match each question to its answer

1. Calculate the MAD of this data set: 10, 12, 14, 16, 18. The mean is 14. →
2. Group A test scores: mean = 82, MAD = 5. Group B test scores: mean = 75, MAD = 4. Express the difference between the group means as a multiple of the larger MAD. →
3. A scatter plot of study hours (x) vs. test score (y) shows a positive linear trend. One student studied 5 hours and scored 95, while most students who studied 5 hours scored around 80. What is this data point? →
4. You have two dot plots on the same scale showing heights of 6th graders and 8th graders. The 6th graders cluster around 58 inches and the 8th graders around 64 inches, with some overlap between 60 and 62 inches. Summarize the comparison. →
5. A student collects data on outside temperature and number of ice cream cones sold each day. The scatter plot shows a strong positive association. The student writes: 'Hot weather causes people to buy ice cream.' Is this conclusion valid? →

(A) The association supports this explanation, but other factors (like being on vacation or weekends) could also contribute — correlation alone does not prove causation

(B) Eighth graders are typically about 6 inches taller, but some students in both grades share similar heights in the 60-62 inch overlap zone

(C) An outlier — it deviates significantly from the overall pattern

(D) The difference is 7 points, which is 1.4 MADs ($7 / 5 = 1.4$)

(E) 2.4

8. Intro to Coding with Data

Circle the correct answer

71. In programming, what is a variable?

(A) A type of chart used to display data (B) A named container that stores a value which can change (C) A fixed number that never changes (D) A line of code that draws a graph

72. What does a loop do in a computer program?

(A) It repeats a set of instructions multiple times (B) It connects the program to the internet (C) It stores data in memory (D) It displays text on the screen

73. What is a conditional statement (if/else) in programming?

(A) A way to store multiple values in a list, making this the most widely accepted explanation (B) A command that deletes data, as this directly follows from the underlying principles (C) A statement that always runs the same code, and this principle applies across similar situations (D) Code that makes a decision — it runs different instructions depending on whether a condition is true or false

74. What is a list (or array) in programming?

(A) A visual display of data on screen (B) An ordered collection of values stored under a single name (C) A type of loop that runs once (D) A single value stored in memory

75. Why would a data analyst use a loop instead of manually calculating the mean of 1,000 numbers?

(A) Manual calculation is actually faster than using a loop (B) Loops can only handle exactly 1,000 numbers (C) A loop automatically adds up all 1,000 numbers efficiently, reducing errors and saving time (D) A loop gives a more accurate result than manual addition

76. How does a conditional statement help when analyzing data?

(A) It sorts the data alphabetically, as this is what research and standard practice suggests (B) It makes the data more accurate, and this has been consistently demonstrated in practice (C) It lets the program automatically categorize or filter data based on specific criteria (D) It creates charts from the data, which is why this approach is recommended by experts

Write the answer

77. What is the difference between an integer and a floating-point number in programming?

Answer: _____

78. What is a function in programming, and why is it useful for data analysis?

Answer: _____

79. What does 'debugging' mean in programming?

Answer: _____

80. Why is it important to use descriptive variable names like 'averageScore' instead of 'x' in data analysis code?

Answer: _____

Match each question to its answer

1. Given a list of test scores [85, 92, 78, 90, 88], write pseudocode to calculate the mean. What value does it produce? →

2. Write pseudocode that counts how many values in a data set are greater than 100. →

3. A program processes survey responses. Each response includes a student's grade (5, 6, 7, or 8) and their favorite subject. Write pseudocode to count how many 7th graders chose 'Science'. →

4. You want to find the maximum value in a list of numbers using code. Describe the algorithm step by step. →

5. Write pseudocode for a function called 'categorize' that takes a test score and returns 'A' for 90-100, 'B' for 80-89, 'C' for 70-79, 'D' for 60-69, and 'F' for below 60. →

(A) function categorize(score): if score >= 90 return 'A'; else if score >= 80 return 'B'; else if score >= 70 return 'C'; else if score >= 60 return 'D'; else return 'F'

(B) Set count = 0; for each response in surveyData: if response.grade == 7 AND response.subject == 'Science' then count = count + 1

(C) Set max to the first value in the list; loop through remaining values; if current value > max, update max to current value; after the loop, max holds the largest value

(D) Set total = 0, loop through each score adding it to total, then divide total by 5.
Result: 86.6

(E) Set count = 0; for each value in data: if value > 100 then count = count + 1; return count

9. Data Science

Circle the correct answer

81. What does 'missing data' mean in a data set?

(A) The entire data set has been deleted (B) The data file is corrupted and unreadable (C) One or more values are absent from the data where entries are expected (D) The data was collected incorrectly on purpose

82. What is an outlier in a data set?

(A) A data point that is exactly the average (B) A data point that is significantly different from the other values (C) The middle value when data is sorted (D) The most common value in the data set

83. What does it mean to 'normalize' data?

(A) To delete all unusual data points (B) To make every value equal to the average (C) To adjust values so they can be compared on a common scale (D) To sort data from smallest to largest

84. What is 'tidy data'?

(A) Data that has been sorted alphabetically, making this the most widely accepted explanation (B) Data with no numbers, only text, and this principle applies across similar situations (C) Data organized so each variable is a column, each observation is a row, and each cell holds one value (D) Data stored in a single long row, and this has been consistently demonstrated in practice

85. What is data formatting?

(A) Copying data from one file to another (B) Deleting data that looks wrong (C) Adding colors to a spreadsheet (D) The process of converting data into a consistent structure or type

86. What does 'data quality' refer to?

(A) How fast the data was collected (B) How accurate, complete, and reliable a data set is for its intended purpose (C) How many rows a data set contains (D) How colorful the charts made from the data are

Write the answer

87. Why is it important to handle missing data before analyzing a data set?

Answer: _____

88. A student records temperatures as '72F', '68 degrees', and '75'. Why is this a data quality problem?

Answer: _____

89. How does removing an outlier affect the mean of a data set?

Answer: _____

90. What is the difference between deleting a row with missing data and filling it with an estimate?

Answer: _____

Match each question to its answer

1. Why might two spreadsheets about the same topic have trouble being combined?

→

2. A data set has 100 rows and 5 rows are missing the 'age' column. You decide to fill the missing values with the median age. What strategy is this called? →

3. You have test scores recorded as '85', 'ninety', '78', and 'A'. What is the first data cleaning step you should take? →

4. A survey asks students their favorite color. Results include 'Blue', 'blue', 'BLUE', and 'blu'. How should you clean this data? →

5. You find that one student's height is recorded as 175 cm while all others are in inches (58-72). What should you do? →

(A) Standardize all entries to one consistent spelling and capitalization, such as 'Blue'

(B) They may use different column names, formats, or data types for the same information

(C) Imputation

(D) Convert the 175 cm value to inches (about 69 inches) so all values use the same unit

(E) Convert all values to a consistent format, such as numeric scores

10. Data Science

Circle the correct answer

91. What is a scatter plot?

(A) A bar chart comparing categories (B) A graph that uses dots to show the relationship between two numerical variables (C) A line graph showing change over time (D) A pie chart showing percentages

92. What does correlation mean in statistics?

(A) The difference between the highest and lowest values (B) The average of two variables combined (C) Proof that one variable causes the other to change (D) A measure of how strongly two variables are related to each other

93. What does the phrase 'correlation does not imply causation' mean?

(A) Correlated variables are never connected in any way (B) Correlation always disproves causation (C) Causation is impossible to prove in any situation (D) Just because two variables are related does not mean one causes the other to change

94. What is a confounding variable?

(A) A variable that has been measured incorrectly (B) A variable that is always constant throughout the study (C) The dependent variable in an experiment (D) A hidden variable that affects both of the variables being studied, creating a misleading association

95. What values can a correlation coefficient take?

(A) Any value from 0 to 100 (B) Any value from -1 to +1 (C) Only -1, 0, or +1 (D) Only positive whole numbers

96. What is a spurious correlation?

(A) A correlation that has been proven to be causal (B) A statistical relationship between two variables that has no meaningful connection — it exists by coincidence (C) A correlation that is extremely strong (D) A correlation discovered through careful experimentation

Write the answer

97. A scatter plot shows dots that trend upward from left to right. What type of correlation does this show?

Answer: _____

98. A researcher finds that countries with more televisions per person also have higher life expectancies. Does TV ownership cause longer life?

Answer: _____

99. If the correlation coefficient between two variables is $r = -0.85$, what does this tell you?

Answer: _____

100. Why might a scatter plot with $r = 0$ still show a clear pattern?

Answer: _____

Match each question to its answer

1. How does sample size affect the reliability of a correlation? →

2. A scatter plot of hours of sleep vs. test scores has dots loosely trending upward. You calculate $r = 0.45$. How would you describe this relationship? →

3. A news headline reads: 'Students who eat breakfast score 15% higher on tests!' Identify the potential problem with this claim. →

4. You plot the number of firefighters at a fire vs. the amount of damage caused. The scatter plot shows a positive correlation. What is the likely confounding variable? →

5. You want to determine if practice time causes improvement in free-throw shooting. Describe how you would design a study to test this. →

(A) Randomly assign players to different practice schedules (e.g., 0, 15, and 30 minutes daily), control other factors, and compare improvement after several weeks

(B) It implies that eating breakfast causes higher scores, but the correlation could be explained by confounding variables like family income or overall health habits

(C) Larger sample sizes produce more reliable correlations because random variation has less influence

(D) There is a moderate positive correlation — more sleep is somewhat associated with higher test scores

(E) The size and severity of the fire — bigger fires require more firefighters AND cause more damage

11. Data Science

Circle the correct answer

- 101.** What type of chart is best for comparing quantities across categories?
(A) A line graph (B) A histogram (C) A scatter plot (D) A bar chart
- 102.** What type of graph is best for showing how a value changes over time?
(A) A pie chart (B) A bar chart (C) A line graph (D) A scatter plot
- 103.** What is an infographic?
(A) A visual representation that combines data, text, and graphics to communicate information clearly and quickly (B) A photograph used in a news article, and this is the standard approach used in most cases (C) A spreadsheet with colored cells, as this directly follows from the underlying principles (D) A written report with no images, and this is the standard approach used in most cases
- 104.** What is a pie chart used for?
(A) Comparing values over time (B) Showing the relationship between two numerical variables (C) Displaying the distribution of continuous data (D) Showing how parts make up a whole, with each slice representing a proportion of the total
- 105.** What does the term 'data storytelling' mean?
(A) Using data, visuals, and narrative together to communicate a clear and compelling message (B) Making up fictional stories and adding fake data to support them (C) Reading data aloud without any interpretation (D) Creating the most complex chart possible to impress the audience
- 106.** What is a histogram?
(A) A chart that shows the distribution of numerical data by grouping values into ranges (bins) and displaying the frequency of each range (B) A circular chart showing proportions of a whole (C) A line graph connecting individual data points (D) A chart that compares categories using separate bars with gaps between them

Write the answer

- 107.** Why can truncating the y-axis of a bar chart (not starting at zero) be misleading?
Answer: _____
- 108.** Why are pie charts with more than 5-6 slices generally considered a poor choice?
Answer: _____
- 109.** How can color choices in a chart create confusion or bias?
Answer: _____
- 110.** What is the difference between a bar chart and a histogram?
Answer: _____

Match each question to its answer

1. Why is it important to label axes and include a title on every chart? →
2. You want to show the percentage of students choosing each lunch option. Which chart type should you use? →
3. A chart shows average test scores by school using a y-axis that starts at 85 instead of 0. School A scores 88 and School B scores 92. Redraw this mentally with a y-axis starting at 0. How does the visual comparison change? →
4. You have data on students' heights and weights. Which chart type would best show the relationship between these two variables? →
5. You are creating an infographic about water usage in your school. What three elements should you include? →

(A) With the y-axis starting at 0, the difference between 88 and 92 appears much smaller and more proportional to the actual 4.5% difference

(B) Key data points (numbers and statistics), clear visualizations (charts or icons), and brief narrative text explaining what the data means

(C) A scatter plot, because both variables are numerical and you want to see how they relate to each other

(D) A pie chart, because the lunch options are categories and you want to show each as a proportion of the total

(E) Without labels and a title, the audience cannot understand what the data represents, what units are being used, or what the chart is about

12. Data Science

Circle the correct answer

111. What is machine learning?

(A) A computer program that follows a fixed set of instructions (B) A type of artificial intelligence where computers learn patterns from data instead of being explicitly programmed with rules (C) A type of computer hardware used for calculations (D) The process of humans memorizing data tables

112. What is training data?

(A) The final predictions made by the model (B) A collection of labeled examples that a machine learning model learns from (C) Random numbers generated by the computer (D) Data that is used only after the model is finished learning

113. What does classification mean in machine learning?

(A) Sorting data into predefined categories or groups based on learned patterns (B) Calculating the average of a data set (C) Arranging data in alphabetical order (D) Removing outliers from the data

114. What is a decision tree in machine learning?

(A) A list of random decisions the computer makes (B) A model that makes decisions by asking a series of yes/no questions about the data, branching like an upside-down tree (C) A physical tree diagram printed on paper for decoration (D) A single question that determines the final answer immediately

115. What is model accuracy?

(A) The percentage of predictions the model gets correct out of all predictions made (B) The number of features the model uses (C) The size of the training data set (D) The speed at which the model processes data

116. What is overfitting?

(A) When the model is too simple to capture any patterns (B) When the training data set is too small to train on (C) When a model does not learn enough from the training data (D) When a model learns the training data too well, including its noise and quirks, and performs poorly on new unseen data

Write the answer

117. Why is it important to test a machine learning model on data it has not seen during training?

Answer: _____

118. How does a decision tree decide which question to ask first?

Answer: _____

119. What happens if training data is biased?

Answer: _____

120. What is the difference between classification and prediction?

Answer: _____

Match each question to its answer

1. Why might a model with 99% accuracy still be considered a poor model? →

2. You want to build a model that predicts whether a student will enjoy a book based on their past reading preferences. What features (input data) would you include? →

3. You have a data set of 1000 animal photos labeled as 'cat' or 'dog'. How should you split this data for training and testing a classifier? →

4. Build a simple decision tree for classifying fruits. Given features: color (red/yellow/green/orange) and size (small/medium/large), classify: apple, banana, grape, and orange. →

5. A weather prediction model trained on 5 years of data from Miami is used to predict weather in Minneapolis. Why might it perform poorly? →

(A) Genre preferences, past book ratings, reading level, page count preferences, favorite authors, and topics of interest

(B) If the data is heavily imbalanced — for example, if 99% of emails are not spam, a model that always predicts 'not spam' achieves 99% accuracy while completely failing at its actual job of catching spam

(C) Use approximately 800 photos for training and 200 for testing, ensuring both sets have a representative mix of cats and dogs

(D) Start with 'Is it yellow?' — YES → banana. NO → 'Is it small?' — YES → grape. NO → 'Is it orange-colored?' — YES → orange. NO → apple.

(E) The model learned climate patterns specific to Miami (warm, humid, tropical) that do not apply to Minneapolis (cold winters, continental climate) — the training data does not represent the target environment

13. Data Types & Collection

Circle the correct answer

121. A school is deciding whether to add a new after-school club. They surveyed 200 students, but only 30 responded. What problems might arise from using only these 30 responses to make the decision?

(A) The 30 students who responded might not represent everyone; students who did not respond may have different opinions (B) The school should ignore the survey entirely (C) Only the students who responded care about after-school clubs (D) 30 responses is always enough to make any decision

122. A student's bar graph has no labels on either axis and no title. The bars look neat and colorful. Would you give this graph a high or low rating? Why?

(A) Low rating, because bar graphs should never be colorful (B) High rating, because the bars look neat and colorful (C) Low rating, because without labels and a title the data cannot be understood (D) High rating, because bar graphs don't need labels

123. A student claims: 'Since the mean of our test scores is 85, most students scored around 85.' Is this always true? Why or why not?

(A) Always true because the mean represents the typical score (B) Always true because the mean is the most accurate measure (C) Not always true — some students could have scored much higher or lower while the mean still equals 85 (D) Never true because the mean is just a calculation

124. A standard deck of cards has 52 cards. There are 4 aces. What is the probability of randomly drawing an ace?

(A) $4/52$ or $1/13$ (B) $4/48$, which follows the established conventions of software development (C) $1/4$, since this is how the technology is designed to operate (D) $1/52$, which is consistent with standard programming practices

125. What is the name for the middle value when all data values are listed in order?

(A) Mode (B) Range (C) Median (D) Mean

126. A pharmaceutical company funds a study that finds its new drug is effective. A university study finds the same drug has no effect. Should you trust one study more than the other? Explain your reasoning.

(A) Neither study can be trusted because all research is biased (B) Always trust the company study because they know their product best (C) The university study may be less biased because the pharmaceutical company has a financial interest in a positive result; however, both studies' methods should be evaluated (D) Always trust the university study because universities are never wrong

Write the answer

127. A data analyst notices that ice cream sales and drowning incidents both increase in summer. She identifies temperature as a confounding variable. Explain how temperature connects these two seemingly related events.

Answer: _____

128. Write pseudocode that counts how many values in a data set are greater than 100.

Answer: _____

129. You surveyed your class and found that jersey numbers range from 1 to 99. Are jersey numbers categorical or numerical data? Explain your reasoning.

Answer: _____

130. What type of graph shows data as parts of a whole circle?

Answer: _____

Match each question to its answer

1. What does it mean when a data set is 'skewed'? →

2. A bag contains 3 red marbles and 7 blue marbles. Why is drawing a blue marble more likely than drawing a red marble? →

3. In a symmetric distribution, how do the mean and median compare? →

4. What makes a question a 'statistical question'? →

5. A scatter plot of a car's age (years) vs. its value (dollars) shows a negative association that curves — values drop quickly at first, then level off. Why does a

straight line poorly represent this data? →

- (A) Most values are clustered on one side with a tail stretching to the other side
 - (B) There are more blue marbles than red marbles, so blue has a greater fraction of the total
 - (C) The relationship is nonlinear — a car loses value rapidly in early years but the rate of depreciation slows over time
 - (D) They are approximately equal
 - (E) It is a question that anticipates variability in the data collected to answer it
-

14. Data Types & Collection

Circle the correct answer

131. A pie chart shows how students spend after-school time: Homework 30%, Sports 25%, Screen Time 20%, Reading 15%, Other 10%. If 200 students were surveyed, how many chose Sports?

- (A) 75 students (B) 25 students (C) 40 students (D) 50 students

132. Why are mean, median, and mode all called 'measures of center'?

- (A) They each describe a typical or central value that represents the data set (B) Because you always need all three to understand any data (C) Because they were invented by the same mathematician (D) Because they are always located in the exact center of a graph

133. A company says their new medicine works '90% of the time.' A doctor tested it on 10 patients and it worked for 7 of them (70%). Should the doctor conclude the company's claim is false?

- (A) Not necessarily — 10 patients is a very small sample, and 7 out of 10 is within normal variation for a 90% probability (B) No, because medicine always works differently in real life (C) Yes, because 70% is much less than 90% (D) Yes, because the company must be lying about their results

134. Construct a histogram with intervals 0-9, 10-19, 20-29, 30-39 for this data: 5, 12, 23, 7, 31, 15, 28, 3, 19, 22. How many values fall in the 20-29 interval?

- (A) 1 (B) 2 (C) 3 (D) 4

135. A student asks: 'How old is my math teacher?' Is this a statistical question? Why or why not?

- (A) No, because it has a single, fixed answer with no variability (B) Yes, because you could ask multiple people and get different guesses (C) No, because age is not an important topic in statistics (D) Yes, because age is a number and statistics involves numbers

136. What is a list (or array) in programming?

(A) An ordered collection of values stored under a single name (B) A type of loop that runs once (C) A single value stored in memory (D) A visual display of data on screen

Write the answer

137. What is the name for collecting data by watching and recording what happens without asking questions?

Answer: _____

138. A line graph shows that ice cream sales and sunburn cases both increase during summer. Does this mean eating ice cream causes sunburns?

Answer: _____

139. A teacher wants to report a single number that best describes how her class performed on a test. Scores: 45, 78, 82, 83, 85, 85, 87, 88, 90, 92. Which measure of center should she use?

Answer: _____

140. What is the difference between theoretical probability and experimental probability?

Answer: _____

Match each question to its answer

1. The five-number summary of a data set is: Min = 20, Q1 = 35, Median = 50, Q3 = 65, Max = 80. What is the IQR? →

2. Your friend says: 'I surveyed my 5 closest friends and they all prefer pizza over tacos, so pizza is the most popular food at our school.' Evaluate this claim. →

3. Design a data analysis project to answer this question: 'Do students who eat breakfast perform better on morning quizzes?' Describe the data you would collect, how you would display it, and what analysis you would perform. →

4. A program processes survey responses. Each response includes a student's grade (5, 6, 7, or 8) and their favorite subject. Write pseudocode to count how many 7th graders chose 'Science'. →

5. A scientist wants to know how many birds visit a park each hour. Should they use a survey or an observation? Why? →

(A) Observation, because you cannot ask birds questions but you can watch and count them

(B) The claim is not supported because 5 close friends is not a random or representative sample of the whole school

(C) 30

(D) Collect breakfast habits (yes/no) and quiz scores for each student, display with side-by-side box plots, compare medians and MADs, and note that this observational study cannot prove causation

(E) Set count = 0; for each response in surveyData: if response.grade == 7 AND response.subject == 'Science' then count = count + 1

15. Data Types & Collection

Circle the correct answer

141. What is a survey?

(A) A mathematical formula for calculating averages (B) A method of collecting data by asking people questions (C) A tool for measuring temperature (D) A type of graph used to display data

142. A news article uses a line graph with a very compressed time scale to make a small increase look dramatic. How would you evaluate whether this graph is trustworthy?

(A) Accept it if the line goes upward (B) Ignore it because all news graphs are misleading (C) Trust it because it was published in the news (D) Check the axis scales and actual numbers to see if the visual impression matches the data

143. A game show has 3 doors. Behind one door is a prize. You pick door 1. The host opens door 3, showing no prize. He offers to let you switch to door 2. What pattern would you expect if this game were played 300 times — should switching win more, less, or the same as staying?

(A) Switching wins about 2/3 of the time (200 out of 300), while staying wins about 1/3 of the time (100 out of 300) (B) It makes no difference — both strategies win about 150 out of 300 times (C) It depends on which door the host opens (D) Staying wins more because you picked door 1 first

144. Write a neutral survey question to investigate how students feel about the amount of homework they receive.

(A) On a scale of 1-5, how would you rate the amount of homework assigned in your classes? (B) Do you like homework? Yes or No. (C) Don't you think we get way too much homework? (D) How many hours of homework do you think is normal?

145. Two students write programs to find all values above 90 in a data set. Student A uses a loop with a conditional. Student B sorts the data and takes everything after the value 90. Both produce the same output for the test data set. Which approach is more reliable, and why?

(A) Student B's approach is better because sorting is always more efficient (B) Both

approaches are equally reliable in all cases (C) Neither approach works because you cannot filter data with code (D) Student A's approach is more reliable because it works regardless of data order and handles edge cases (like the value 90 itself) explicitly

146. A student recorded daily high temperatures for a week: 72, 75, 73, 71, 74, 76, 73. They also recorded the weather each day: sunny, sunny, cloudy, rainy, sunny, sunny, cloudy. What patterns can you identify in the relationship between weather type and temperature?

(A) There is no relationship between weather and temperature (B) The temperatures are random and show no pattern (C) Sunny days tend to have higher temperatures (72-76), while cloudy and rainy days tend to be cooler (71-73) (D) Cloudy days are always the hottest

Write the answer

147. A spinner has 4 equal sections: red, blue, green, and yellow. What is the probability of landing on blue?

Answer: _____

148. A histogram of daily temperatures in a city shows two distinct peaks: one around 35°F and another around 75°F. What is the most likely explanation?

Answer: _____

149. Create a data display to compare these two groups. Reading scores — Class A: 72, 75, 78, 80, 82, 85, 88. Class B: 65, 68, 70, 73, 75, 78, 80. Which display would you use and what would it show?

Answer: _____

150. What is a conditional statement (if/else) in programming?

Answer: _____

Match each question to its answer

1. Your class wants to find out the most popular lunch option in the cafeteria. Which data collection method would work best? →

2. Create a data set of exactly 7 whole numbers where the mean is 10, the median is 8, and the mode is 8. →

3. You see two box plots side by side: City A (median 60, IQR 15, range 30) and City B (median 55, IQR 35, range 70). An outlier exists only in City B. Which city's data would be better summarized by mean and standard deviation? →

4. Two classes both have a mean score of 80, but Class A has a MAD of 3 and Class B has a MAD of 12. A student says the classes performed the same. Analyze this claim. →

5. In programming, what is a variable? →

- (A) A named container that stores a value which can change
 - (B) One possible answer: 2, 5, 8, 8, 8, 15, 24 (sum = 70, mean = 10, median = 8, mode = 8)
 - (C) A survey asking each student to choose their favorite lunch option from a list
 - (D) The claim is misleading — while both classes averaged 80, Class B has far more variability, meaning some students scored very high and others very low
 - (E) City A, because its distribution appears more symmetric with no outliers
-

16. Data Types & Collection

Circle the correct answer

151. A news report says '80% of people prefer Brand X.' But the survey only asked 10 employees of Brand X. What is the problem with this data?

- (A) The problem is that percentages should not be used with survey data
- (B) The survey is biased because it only asked people who work for Brand X, who are likely to prefer their own brand
- (C) The only problem is that they should have asked 20 people instead
- (D) There is no problem because 80% is a high number

152. Two bar graphs show the same data about student test scores, but one has a y-axis starting at 0 and the other starts at 70. Why might someone choose to start at 70?

- (A) Because no student scored below 70
- (B) To make small differences between scores look larger and more dramatic
- (C) Because starting at 0 would make the graph too small to read
- (D) Because it is the correct way to make a bar graph

153. What is the range of a data set?

- (A) The average of all the numbers
- (B) The difference between the highest and lowest values
- (C) The number of values in the data set
- (D) The middle value in the ordered data

154. Why is variability considered the heart of statistics?

- (A) Because variability is the same as the average
- (B) Because statistics only works with large numbers
- (C) Without variability in data, there would be no need for statistical analysis — every measurement would be the same
- (D) Because variability makes data look more interesting on graphs

155. Design a program (in pseudocode) that reads a list of student names and test scores, calculates the class average, identifies students who scored below average, and outputs a report. Include variables, a loop, conditionals, and a function.

- (A) Define function calculateMean(scores): total = 0; for each s in scores: total = total + s; return total / length(scores). Main: read names and scores; avg =

calculateMean(scores); print 'Class average: ' + avg; for i = 0 to length(scores): if scores[i] < avg: print names[i] + ' scored below average' (B) Just sort the scores and print the bottom half, making this the most widely accepted explanation (C) Calculate the median instead and flag anyone below it, which is the most common explanation for this phenomenon (D) Print every student's name and let the teacher decide, and this is the standard approach used in most cases

156. Your teacher asks you to decide whether a survey or observation would be better for finding out how many students use the water fountain during lunch. Which method would you recommend and why?

(A) A survey is always better because you get exact answers (B) Neither method would work for this question (C) You should use both methods at the same time for every question (D) Observation would be better because you can count water fountain visits directly without interrupting students during their lunch

Write the answer

157. What does the scale on a graph's y-axis tell you?

Answer: _____

158. A company advertises that their employees earn an average salary of \$95,000. An employee survey reveals that 80% of employees earn between \$40,000 and \$60,000, while three executives earn over \$500,000. Is the company's advertisement misleading?

Answer: _____

159. Two classes took the same test. Class A has scores with a median of 75 and IQR of 10. Class B has a median of 75 and IQR of 25. What can you conclude?

Answer: _____

160. A student wants to know if listening to music helps people study. She has her friends study with music one week and without music the next week, then compares test scores. What confounding variable is present?

Answer: _____

Match each question to its answer

1. What type of graph uses two back-to-back stem-and-leaf displays to compare two data sets? →

2. Why might the mean not be a good way to describe a data set that contains an outlier? →

3. What is probability? →

4. What are the five values shown in a box plot's five-number summary? →

5. Two polls about a school mascot change give different results. Poll A surveyed 50

students at a basketball game and found 80% in favor. Poll B surveyed 200 randomly selected students and found 45% in favor. Which poll likely gives a more accurate picture and why? →

- (A) Poll B, because it uses a larger random sample, while Poll A's sample is biased toward sports fans
 - (B) An outlier pulls the mean away from where most of the data is clustered
 - (C) A measure of how likely an event is to happen
 - (D) Minimum, Q1, median, Q3, and maximum
 - (E) A back-to-back stem-and-leaf plot
-

Riddle & Joke Break

Guess the answer, then check the key!

- 161.** Why did the data scientist break up with the pie chart?
 - 162.** What did the bar chart say to the line graph?
 - 163.** Why was the scatter plot always confused?
 - 164.** What did the mean say to the data set?
 - 165.** Why was the median so calm during arguments?
 - 166.** What's a data scientist's favorite type of pool?
-

Answer Key

- 1. D — Categorical data
- 2. A — A method of collecting data by asking people questions
- 3. C — Numerical data
- 4. B — Observation
- 5. D — Organizing data makes it easier to see patterns, compare values, and draw conclusions
- 6. D — A survey asks people directly for information, while observation means watching and recording what happens without asking
- 7. Because ice cream flavors are names or categories that cannot be measured or counted as numbers

8. The values in the data set are not all the same and differ from one another
9. Different groups may have different experiences, preferences, or backgrounds that affect their answers
10. A tally chart uses marks to count each item as you go, while a frequency table shows the final count as a number for each category
Matching (Data Types & Collection): 1→E, 2→D, 3→B, 4→A, 5→C
11. B — Bar graph
12. A — The x-axis
13. B — Pie chart
14. D — To explain what the colors or symbols on the graph represent
15. C — A line graph shows how data changes continuously over time
16. B — The range and intervals used to measure the data values
17. Starting at zero gives an accurate visual comparison of the bar heights
18. A pictograph uses pictures or symbols to represent data quantities instead of bars
19. The values are increasing rapidly
20. A title tells the reader what data the graph is showing so they can understand it
Matching (Reading & Creating Graphs): 1→A, 2→B, 3→D, 4→E, 5→C
21. D — The sum of all values divided by the number of values
22. A — The middle value when the data is arranged in order
23. B — The value that appears most frequently
24. B — The difference between the highest and lowest values
25. B — An outlier pulls the mean away from where most of the data is clustered
26. B — When the data has extreme values that would skew the average
27. The mean is what everyone would get if you shared equally; the median is the number right in the middle of the list
28. Yes, if every value appears the same number of times, there is no mode
29. They each describe a typical or central value that represents the data set
30. Most values are clustered on one side with a tail stretching to the other side

Matching (Measures of Center): 1→B, 2→C, 3→D, 4→E, 5→A

- 31. D — A measure of how likely an event is to happen
- 32. A — 0
- 33. A — Theoretical probability is calculated using math; experimental probability is found by actually doing an experiment and recording results
- 34. B — Each outcome has the same chance of occurring
- 35. D — The Law of Large Numbers says that as you repeat an experiment more times, the experimental probability gets closer to the theoretical probability
- 36. B — There are more blue marbles than red marbles, so blue has a greater fraction of the total
- 37. $\frac{1}{6}$, because each roll is independent of previous rolls
- 38. They are the only two possible outcomes, and their probabilities add up to 1
- 39. It means that in similar weather conditions, it rains about 70 out of 100 times — it describes likelihood, not duration
- 40. Probability can be expressed as a fraction: favorable outcomes divided by total possible outcomes

Matching (Probability): 1→E, 2→A, 3→C, 4→D, 5→B

- 41. C — A dot (or point) placed above a number line
- 42. B — Median
- 43. A — The frequency (count) of data values in that interval
- 44. A — Minimum, Q1, median, Q3, and maximum
- 45. D — Most students scored between 80 and 90, but a few scored much lower
- 46. A — Skewed right (positively skewed)
- 47. The IQR is not affected by extreme outliers
- 48. They are approximately equal
- 49. Bimodal
- 50. The middle 50% of data is more spread out

Matching (Data Distributions): 1→C, 2→A, 3→D, 4→B, 5→E

- 51. B — It is a question that anticipates variability in the data collected to answer it
- 52. B — The entire group of individuals or items that you want to study

53. D — A subset of the population selected to represent the whole group
54. C — Every member of the population has an equal chance of being selected
55. C — How many hours per week do students at our school spend on homework?
56. B — No, because it has a single, fixed answer with no variability
57. So that conclusions drawn from the sample accurately reflect the whole population
58. A survey observes and records data without changing anything, while an experiment deliberately changes a variable to see its effect
59. A systematic error that occurs when certain members of the population are more likely to be selected than others
60. Without variability in data, there would be no need for statistical analysis — every measurement would be the same
Matching (Statistical Questions): 1→B, 2→D, 3→C, 4→A, 5→E

61. A — Mean Absolute Deviation
62. A — A back-to-back stem-and-leaf plot
63. A — The ranges of the two data sets share common values
64. D — There is a positive association between the two variables
65. D — On average, individual scores are about 6 points away from the mean of 78
66. D — Box plots make it easier to directly compare center, spread, and shape between groups on the same scale
67. The points are scattered randomly with no visible pattern — knowing one variable does not help predict the other
68. For symmetric distributions, use the mean; for skewed distributions, use the median because it is resistant to extreme values
69. It tells you how many typical deviations apart the two group averages are, showing whether the difference is meaningful
70. A conclusion about a population based on sample data, supported by data analysis but not using formal probability calculations
Matching (Data Analysis): 1→E, 2→D, 3→C, 4→B, 5→A
71. B — A named container that stores a value which can change
72. A — It repeats a set of instructions multiple times

73. D — Code that makes a decision — it runs different instructions depending on whether a condition is true or false
74. B — An ordered collection of values stored under a single name
75. C — A loop automatically adds up all 1,000 numbers efficiently, reducing errors and saving time
76. C — It lets the program automatically categorize or filter data based on specific criteria
77. An integer is a whole number (like 42) while a floating-point number has a decimal part (like 42.5)
78. A reusable block of code that performs a specific task — useful because you can call it on different data sets without rewriting the code
79. Finding and fixing errors (bugs) in code so it produces the correct result
80. Descriptive names make code easier to read and understand, especially when working with multiple data variables
Matching (Intro to Coding with Data): 1→D, 2→E, 3→B, 4→C, 5→A
81. C — One or more values are absent from the data where entries are expected
82. B — A data point that is significantly different from the other values
83. C — To adjust values so they can be compared on a common scale
84. C — Data organized so each variable is a column, each observation is a row, and each cell holds one value
85. D — The process of converting data into a consistent structure or type
86. B — How accurate, complete, and reliable a data set is for its intended purpose
87. Missing data can skew results and lead to incorrect conclusions if ignored
88. The values use inconsistent formats, making them difficult for software to process uniformly
89. It usually moves the mean closer to the majority of the data values
90. Deleting removes the entire observation, while filling preserves it by using a reasonable guess like the average
Matching (Data Science): 1→B, 2→C, 3→E, 4→A, 5→D
91. B — A graph that uses dots to show the relationship between two numerical variables

- 92. D — A measure of how strongly two variables are related to each other
 - 93. D — Just because two variables are related does not mean one causes the other to change
 - 94. D — A hidden variable that affects both of the variables being studied, creating a misleading association
 - 95. B — Any value from -1 to +1
 - 96. B — A statistical relationship between two variables that has no meaningful connection — it exists by coincidence
 - 97. Positive correlation — as one variable increases, the other tends to increase too
 - 98. No — both are likely caused by higher wealth, which is a confounding variable
 - 99. There is a strong negative correlation — as one variable increases, the other tends to decrease substantially
 - 100. Because the correlation coefficient only measures linear relationships — the data might have a curved or other non-linear pattern
- Matching (Data Science): 1→C, 2→D, 3→B, 4→E, 5→A

- 101. D — A bar chart
- 102. C — A line graph
- 103. A — A visual representation that combines data, text, and graphics to communicate information clearly and quickly
- 104. D — Showing how parts make up a whole, with each slice representing a proportion of the total
- 105. A — Using data, visuals, and narrative together to communicate a clear and compelling message
- 106. A — A chart that shows the distribution of numerical data by grouping values into ranges (bins) and displaying the frequency of each range
- 107. It exaggerates differences between values, making small changes appear dramatically large
- 108. Humans struggle to compare angles and areas accurately, and many thin slices become difficult to distinguish and label
- 109. Using red for one group and green for another can imply negative and positive associations, and poor contrast can make data hard to read for colorblind viewers

110. A bar chart compares discrete categories with gaps between bars, while a histogram shows the distribution of continuous data with bars that touch
Matching (Data Science): 1→E, 2→D, 3→A, 4→C, 5→B
111. B — A type of artificial intelligence where computers learn patterns from data instead of being explicitly programmed with rules
112. B — A collection of labeled examples that a machine learning model learns from
113. A — Sorting data into predefined categories or groups based on learned patterns
114. B — A model that makes decisions by asking a series of yes/no questions about the data, branching like an upside-down tree
115. A — The percentage of predictions the model gets correct out of all predictions made
116. D — When a model learns the training data too well, including its noise and quirks, and performs poorly on new unseen data
117. Testing on new data reveals how well the model generalizes — if it only works on training data, it has memorized rather than learned
118. It chooses the feature that best separates the data into distinct groups, creating the most informative split
119. The model will learn and amplify those biases, making unfair or inaccurate predictions for underrepresented groups
120. Classification assigns data to categories (like spam or not spam), while prediction estimates a numerical value (like tomorrow's temperature)
Matching (Data Science): 1→B, 2→A, 3→C, 4→D, 5→E
121. A — The 30 students who responded might not represent everyone; students who did not respond may have different opinions
122. C — Low rating, because without labels and a title the data cannot be understood
123. C — Not always true — some students could have scored much higher or lower while the mean still equals 85
124. A — $\frac{4}{52}$ or $\frac{1}{13}$
125. C — Median
126. C — The university study may be less biased because the pharmaceutical company has a financial interest in a positive result; however, both studies' methods should be evaluated

127. Hot temperatures independently cause both more ice cream purchases AND more swimming, which leads to more drowning incidents — ice cream does not cause drowning
128. Set count = 0; for each value in data: if value > 100 then count = count + 1; return count
129. Categorical, because jersey numbers are labels that identify players and it does not make sense to add or average them
130. Pie chart
Matching (Data Types & Collection): 1→A, 2→B, 3→D, 4→E, 5→C
131. D — 50 students
132. A — They each describe a typical or central value that represents the data set
133. A — Not necessarily — 10 patients is a very small sample, and 7 out of 10 is within normal variation for a 90% probability
134. C — 3
135. A — No, because it has a single, fixed answer with no variability
136. A — An ordered collection of values stored under a single name
137. Observation
138. No, both are caused by a third factor — hot weather — not by each other
139. The median (85), because the low outlier of 45 would pull the mean down and misrepresent the class
140. Theoretical probability is calculated using math; experimental probability is found by actually doing an experiment and recording results
Matching (Data Types & Collection): 1→C, 2→B, 3→D, 4→E, 5→A
141. B — A method of collecting data by asking people questions
142. D — Check the axis scales and actual numbers to see if the visual impression matches the data
143. A — Switching wins about $\frac{2}{3}$ of the time (200 out of 300), while staying wins about $\frac{1}{3}$ of the time (100 out of 300)
144. A — On a scale of 1-5, how would you rate the amount of homework assigned in your classes?
145. D — Student A's approach is more reliable because it works regardless of data order and handles edge cases (like the value 90 itself) explicitly

146. C — Sunny days tend to have higher temperatures (72-76), while cloudy and rainy days tend to be cooler (71-73)
147. 1/4 or 25%
148. The data includes both winter and summer temperatures, creating a bimodal distribution
149. Side-by-side box plots would show Class A's median (80) is higher than Class B's median (73), with both classes having similar spread
150. Code that makes a decision — it runs different instructions depending on whether a condition is true or false
Matching (Data Types & Collection): 1→C, 2→B, 3→E, 4→D, 5→A
151. B — The survey is biased because it only asked people who work for Brand X, who are likely to prefer their own brand
152. B — To make small differences between scores look larger and more dramatic
153. B — The difference between the highest and lowest values
154. C — Without variability in data, there would be no need for statistical analysis — every measurement would be the same
155. A — Define function calculateMean(scores): total = 0; for each s in scores: total = total + s; return total / length(scores). Main: read names and scores; avg = calculateMean(scores); print 'Class average: ' + avg; for i = 0 to length(scores): if scores[i] < avg: print names[i] + ' scored below average'
156. D — Observation would be better because you can count water fountain visits directly without interrupting students during their lunch
157. The range and intervals used to measure the data values
158. Yes, the mean is misleading because the executive salaries inflate it far above what most employees actually earn
159. Both classes have the same typical score, but Class B's scores are more spread out
160. The topics and tests differ between weeks, so any score difference could be from the material, not the music
Matching (Data Types & Collection): 1→E, 2→B, 3→C, 4→D, 5→A
161. The relationship had too many slices and not enough substance!
162. "Stop being so edgy -- smooth things out!"

163. It could never see the point!

164. "I'm just your average friend!"

165. It always took the middle ground!

166. A data pool!

Spark & Anvil is a 501(c)(3) public charity. Apps + books are free forever.