Name: _______________ **Date:** _______________

Answer Key: Data Integrity & Algorithmic Bias: 9th Grade Cybersecurity Literacy Quiz

High schoolers identify structural bias in machine learning and apply data cleaning techniques to maintain digital integrity and ethical accuracy.

---

1. An urban planning group uses transit app data to determine where to build new bike lanes. What is the most significant 'sampling bias' risk in this dataset?

Answer: A) The data only represents residents who own smartphones and use that specific app.

Sampling bias occurs when the collected data is not representative of the entire population; in this case, it excludes people without smartphones or those who don't use the app.

2. Data 'scrubbing' or cleaning is the process of removing outliers or errors from a dataset to improve the accuracy of the final analysis.

Answer: A) True

Data cleaning/scrubbing is a critical step in data literacy to ensure that 'dirty data' (duplicates, errors, or outliers) does not lead to incorrect conclusions.

3. When a researcher uses data collected by a government agency (like the World Health Organization) rather than gathering it themselves, they are using ______ data.

Answer: B) Secondary

Secondary data is information that has already been collected and processed by others, which requires users to evaluate the original collector's methodology.

4. Which of these is a 'proxy variable' for measuring a person's socioeconomic status if direct income data is unavailable?

Answer: C) Their highest level of education completed

A proxy variable is an indirect measure of a value; education level often correlates strongly with socioeconomic status when direct data is missing.

5. Correlation between two data points (such as ice cream sales and sunburns) always proves that one variable caused the other to happen.

Answer: B) False

Correlation does not equal causation; a third variable (like hot weather) often causes both, making the direct link between the two points a logical fallacy.

6. The ethical practice of making data 'anonymous' by removing names, Social Security numbers, and birthdates is known as ______.

Answer: A) De-identification

De-identification is a key component of data management and privacy, ensuring that individuals cannot be identified from a dataset.

7. A data visualization uses a truncated y-axis (starting at 50 instead of 0) to show a small increase in stock prices. Why might this be considered misleading?

Answer: B) It makes a small change look much more dramatic than it actually is.

Manipulating the scale of an axis can exaggerate trends, which is a common way data is used to mislead audiences in media and advertising.

8. What is the primary risk of using 'Low-Quality' data (data that is outdated or inaccurate) in an Artificial Intelligence model?

Answer: C) The model will produce 'Garbage In, Garbage Out' (GIGO) results.

GIGO is a fundamental concept in CS; if the input data is flawed, the output generated by the algorithm or AI will also be flawed and unreliable.

9. When evaluating a source, a data scientist looks at the ______, which describes the history, origins, and movements of a dataset.

Answer: A) Data Provenance

Data provenance (or lineage) allows researchers to trace data back to its source to ensure it hasn't been tampered with or misinterpreted over time.

10. Open Data initiatives are projects that make government and scientific datasets freely available for anyone to use, redistribute, and reuse.

Answer: A) True

Name: ________________　　**Date: ________________**

Open Data is a movement to increase transparency and innovation by allowing public access to important datasets like weather, traffic, and health statistics.