Name: _____   **Date: _____**

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?

    A.   The data only represents residents who own smartphones and use that specific app.
    B.   The data is stored in a cloud-based server rather than a local hard drive.
    C.   The app uses GPS coordinates instead of manual address entries.
    D.   The dataset includes information from both weekends and weekdays.

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

    A.   True
    B.   False

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.

    A.   Primary
    B.   Secondary
    C.   Qualitative
    D.   Anecdotal

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

    A.   Their favorite color
    B.   The number of hours they sleep
    C.   Their highest level of education completed
    D.   The operating system of their computer

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

    A.   True
    B.   False

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

    A.   De-identification
    B.   Categorization
    C.   Data Mining

    D. Encryption

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?

    A. It hides the fact that the data is qualitative.

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

    C. It prevents the reader from seeing the historical data.

    D. It identifies the source of the data too clearly.

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 What is the primary risk of using 'Low-Quality' data (data that is outdated or inaccurate) in an Artificial Intelligence model?

    A. The model will take up too much storage space.

    B. The model will automatically delete the data.

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

    D. The model will require more electricity to run.

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

    A. Data Provenance

    B. Data Volume

    C. File Extension

    D. Metadata Tagging

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

    A. True

    B. False