Eugene Fotsing

CTEC 298

**Understanding Data: A Foundation for Data Science**

A dataset is a structured collection of information, typically organized into elements, variables, and observations. **Elements** are the individual entities being studied, such as people, companies, or products. **Variables** are the characteristics or features of these elements that we're interested in measuring, such as age, income, or product sales. **Observations** are the specific values of these variables for each element.

Data can be broadly classified into two types:

1. **Qualitative Data:** This data type represents categories or labels, such as gender, color, or product type. It can be further categorized as:
   * **Nominal Data:** Data without any inherent order, like eye color or country of origin.
   * **Ordinal Data:** Data with a specific order or ranking, such as education level or customer satisfaction ratings.
2. **Quantitative Data:** This type of data represents numerical values. It can be:
   * **Discrete Data:** Data that can only take on specific values, often whole numbers, such as the number of children in a family or the number of products sold.
   * **Continuous Data:** Data that can take on any value within a range, such as height, weight, or temperature.

Understanding the type of data is crucial in data science as it determines the appropriate statistical techniques to use.

**Key Statistical Measures**

To gain insights from data, we often use statistical measures:

* **Measures of Central Tendency:**
  + **Mean:** The average value of a dataset.
  + **Median:** The middle value when data is sorted.
  + **Mode:** The most frequent value in a dataset.
* **Measures of Dispersion:**
  + **Range:** The difference between the highest and lowest values.
  + **Variance:** A measure of how spread out the data is from the mean.
  + **Standard Deviation:** The square root of variance, providing a more interpretable measure of spread.

By understanding these basic statistical concepts, data scientists can uncover patterns, make predictions, and derive valuable insights from datasets. This foundation is essential for a wide range of data-driven applications.

**References:**

1. Wickham, H. (2014). "Tidy Data." *Journal of Statistical Software*.
2. Jones, M. T. (2017). "Cleansing, Processing, and Visualizing a Data Set, Part 1: Working with Messy Data." *IBM DeveloperWorks*.