Course: Data and Information Literacy

Lecture: 3 Answers to Self-Assessment Questions

Lecturer: Dr. Johnson Masinde

1. **What are the key differences between descriptive, inferential, and predictive data analysis methods?**

- **Descriptive Data Analysis** focuses on summarizing and describing the main features of a dataset. It involves statistical measures such as mean, median, mode, and standard deviation to provide a snapshot of the data. It does not make predictions or generalizations beyond the dataset.
- **Inferential Data Analysis** uses a small sample of data to make inferences or generalizations about a larger population. This method involves hypothesis testing, confidence intervals, and regression analysis. The goal is to draw conclusions that extend beyond the immediate data.
- **Predictive Data Analysis** aims to predict future outcomes based on historical data. It leverages techniques such as machine learning models, regression analysis, and time series forecasting. Unlike descriptive and inferential analysis, predictive analysis uses past data patterns to make forecasts about future events.

2. **How does exploratory data analysis (EDA) help in identifying data patterns and anomalies, and what techniques are commonly used during EDA?**

- **Exploratory Data Analysis (EDA)** is a process used to examine datasets to uncover underlying structures, detect outliers or anomalies, and identify relationships among variables. EDA helps in understanding the data distribution and variability before applying any formal modeling.
- Techniques commonly used in EDA include:
    - **Visualization Tools**: Scatter plots, histograms, box plots, and heatmaps are used to visualize the distribution, relationships, and patterns in the data.

- **Summary Statistics**: Calculating mean, median, variance, skewness, and kurtosis to understand the data's central tendency and spread.
- **Outlier Detection**: Identifying extreme values or anomalies through methods like z-scores or interquartile range (IQR).
- **Correlation Analysis**: Checking for linear relationships between variables using correlation matrices or scatter plots.

3. **In the context of machine learning, how do supervised and unsupervised methods differ in their approach to data analysis?**

- **Supervised Learning** uses labeled data to train models, where each input has a corresponding output. The goal is to learn the mapping from inputs to outputs and make predictions. Common algorithms include decision trees, support vector machines (SVMs), and neural networks. In supervised learning, the data analysis focuses on classification (categorizing data) or regression (predicting numerical outcomes).
- **Unsupervised Learning** works with unlabeled data and aims to find hidden patterns or groupings in the data. Since there is no predefined outcome or target variable, the algorithm must discover the structure in the data. Common techniques include clustering (e.g., K-means) and dimensionality reduction (e.g., PCA). In unsupervised learning, data analysis revolves around identifying patterns, segmenting data, or finding anomalies.