UNILAGS Data Science Bootcamp 2024

Data Analytics

Yizhou (Curtis) Chen

STATISTICS WITHOUT BORDERS

Objectives

This course offers an introduction to data analytics, covering the fundamentals of data cleaning and preparation. Participants will explore various techniques and tools, gaining practical skills essential for analyzing and deriving insights from diverse datasets.

Course Outline

- Introduction to Data Analytics
- Understanding Data
- Missingness
- Extreme Values
- Data Analytics
- Exercises

What is Data Analytics?

Definition: Data analytics refers to the science of analyzing raw data to make conclusions about information.

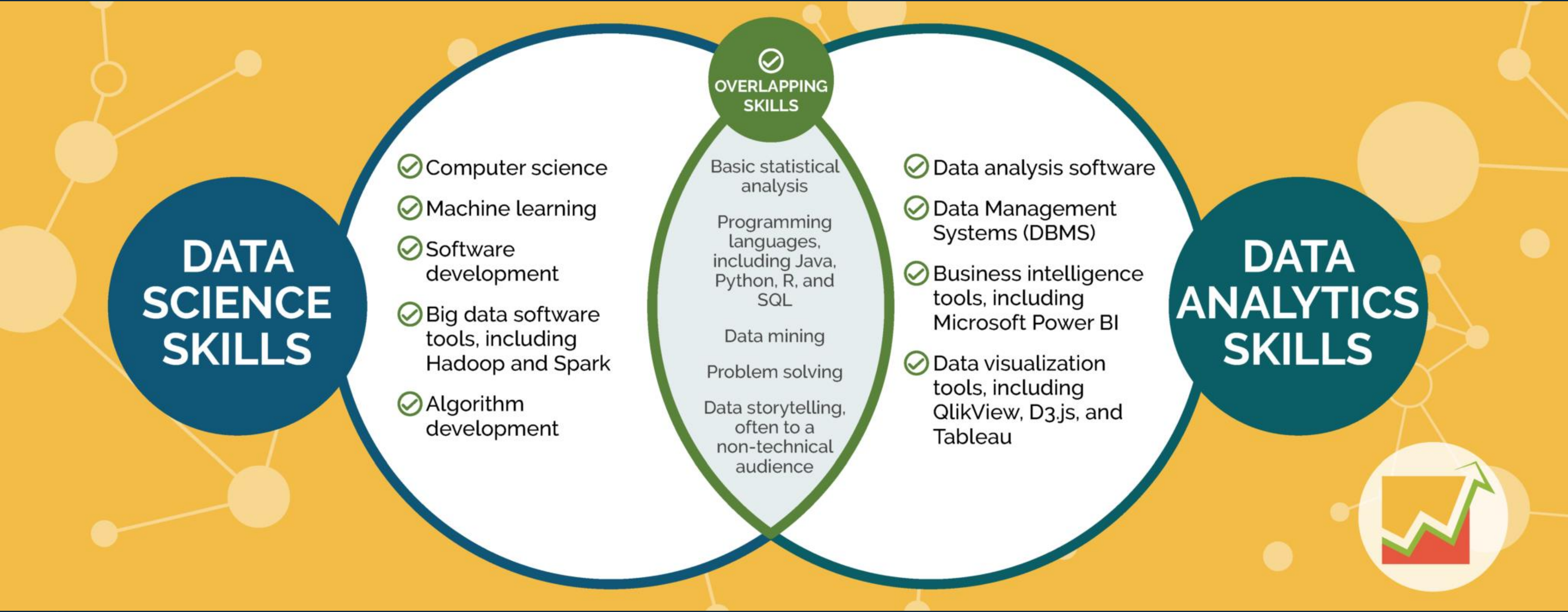
- Using algorithmic or mechanical processes to derive insights and running through various data analysis processes to make it suitable for use.
- Different entities

What is Data Science?

Definition: It is an interdisciplinary academic field that uses statistics, scientific computing, scientific methods, processes, scientific visualization, algorithms and systems to extract or extrapolate knowledge and insights from potentially noisy, structured, or unstructured data.

- Creating Algorithms and Methods

Data Analytics vs Data Science Skills



DATA SCIENCE SKILLS

- ✓ Computer science
- ✓ Machine learning
- ✓ Software development
- ✓ Big data software tools, including Hadoop and Spark
- ✓ Algorithm development

✓ OVERLAPPING SKILLS

- Basic statistical analysis
- Programming languages, including Java, Python, R, and SQL
- Data mining
- Problem solving
- Data storytelling, often to a non-technical audience

DATA ANALYTICS SKILLS

- ✓ Data analysis software
- ✓ Data Management Systems (DBMS)
- ✓ Business intelligence tools, including Microsoft Power BI
- ✓ Data visualization tools, including QlikView, D3.js, and Tableau

Data Analytics vs Data Science

- **Data Science** is a broader field that encompasses various methods and technologies for processing and analyzing data, often focusing on creating new models and algorithms.

- **Data Analytics** is a more focused practice within Data Science, primarily concerned with analyzing existing data to provide actionable insights and inform decision-making.

Example: How Nigeria can tackle outbreak of diseases with data analytics – Interview with Stephanie Omakwu

- “Nigeria faces significant public health challenges, including frequent outbreaks of infectious diseases such as malaria, cholera, Lassa fever, and others.”
- ”Data analytics involves examining raw data to draw conclusions and uncover patterns. When integrated with health data, it can identify trends, correlations, and insights that inform public health decisions,”
- Predictive machine learning, a subset of artificial intelligence, uses algorithms and statistical models to predict future outcomes based on historical data.
- “By learning from past data, these models can forecast disease outbreaks, potential hotspots, and the likely spread of infections,”
- One of the most significant benefits of data analytics and predictive machine learning in public health is the ability to detect outbreaks early.
- “Traditional surveillance systems often rely on reported cases, which can lag behind the actual spread of a disease. Predictive models, however, can analyze real-time data from various sources to identify anomalies that may indicate an emerging outbreak,”
- “Data analytics can optimize resource distribution by pinpointing areas at higher risk of disease outbreaks.”

- Source: Interview by Ayo Onikoyi

Example: Using Machine Learning to Predict Diabetes Mellitus in Nigeria

- Goal: Predict diabetes in people at an early stage.
- Method: Data mining process model of data selection, pre-processing, transformation, data mining and interpretation/evaluation (Sharma & Saxena, 2018).
- The steps included: **data selection**, **transformation of the data using standard scaler**, **splitting the data into 70% for training the models and 30% for testing the models**, using three supervised learning algorithms namely, **KNN, decision trees and ANN**; and lastly **interpretation and evaluation of data** by comparison of the algorithms' respective performance.
- Results: Artificial Neural Networks (ANN) outperformed the K- Nearest Neighbors (KNN) and Decision Tree methods.
- This research contributes to practice by producing a 97.40% accurate and validated ANN model, which can help to predict the future occurrence of diabetes in Nigerians, based on their current health records. This model will help mitigate the diabetes death rate among Nigerians due to late clinical diagnosis.
- This model will be helpful to academicians, policy makers, government officials, medical personnel, and researchers in Nigeria as well as Africa at large.

Source: Predictive Model for Diabetes Mellitus Using Machine Learning Techniques (A Study in Nigeria)," The African Journal of Information Systems: Vol. 15: Iss. 1, Article 1.
Available at: https://digitalcommons.kennesaw.edu/ajis/vol15/iss1

Understanding Data

Structured Data and Unstructured Data

- **Structured Data:** Structured data refers to information that is organized in a highly organized, easily searchable format, typically stored in databases or spreadsheets. This type of data is often arranged in rows and columns, making it simple to retrieve, query, and analyze using standard database tools.

- **Unstructured Data:** Unstructured data refers to information that does not have a pre-defined data model or organization. This data type is often more complex and not easily searchable or analyzable using traditional tools. Unstructured data can come in many formats, including text, images, audio, and video.

Numerical Data

- **Continuous Data:** Data that can take any value within a continuous range.

  Examples: Temperature, income, or distance.

- **Discrete Data:** Data that can only take integer values, usually as a result of counting.

  Examples: Number of employees in a company
  Number of customers in a store
  Number of defective products.

Categorical Data & Ordinal Data

- **Categorical Data:** AKA qualitative data, refers to data that is grouped into categories or classes which are typically nominal or ordinal in nature. These categories represent qualitative attributes and are distinctly separate from each other.

- **Ordinal Data:** This is a special type of categorical data where the categories have a natural order or ranking, but the intervals between the categories are unknown or unquantifiable.

  Examples: Customer satisfaction ratings such as "dissatisfied", "neutral", "satisfied", and "very satisfied".

Time Series Data & Spatial Data

- **Time Series Data:** Data points arranged in order of time, commonly used to analyze trends, seasonal variations, or forecasts.

  Examples:
  Daily sales figures
  Monthly visitor counts
  Annual financial reports

- **Spatial Data:** Data related to geographic locations, commonly used for mapping, location services, and geospatial analysis.

  Examples:
  Geographic coordinates
  Urban planning data
  The location data of mobile devices

Text Data & Multimedia Data

- **Text Data:** Typically refers to data in written language form, which can be unstructured, such as reports, emails, and social media posts.

  Examples: Topic modeling
  Sentiment analysis

- **Multimedia Data:** Includes image, audio, and video data. These types of data usually require complex processing techniques, such as computer vision and natural language processing technologies.

Missing Values

Missing Values

- Identifying missing values
- Missing data may be identified as:
  - Blank space
  - Lack of any value, e.g. in .CSV Data set with variables name, age, gender

    Sally,25, ← No missing data

    Bill,,M ← Age is missing
  - A special value: Asterisk, NA, Null, period are common. Legacy dataset may use 9, 99, or 999.
  - CHECK Database documentation to verify how missing data is recorded.

Missing Values

- Treating missing values
  - **Complete Case Analysis or list wise deletion**: Ignore or delete records with a missing value in any variable
  - **Partial Case Analysis or pairwise deletion**: Ignore or delete records with missing value in variable being analyzed
  - **Estimate missing value**
    - Average or median of entire data set
    - Average or median or relevant subset
    - Regression estimate
    - Hot Deck Imputation (use value of previous record in the data set)

Complete Case Analysis

- Assumes data is missing at random
- Easy to implement
- When to use:
  - Not more than about 5% of records have missing data
  - Large dataset ; missing data is not very informative
- Downside
  - Can lead to biased data
  - Reasons for missing information are not addressed
  - Missing at random assumption is rarely true

Partial Case Analysis

- Assumes data is missing at random
- Easy to implement
- Less data is discarded than in complete case analysis
- When to use:
    - Not more than about 5% of records have missing data
- Downsides
    - Results for subsets of the data are likely to be distorted
    - Can lead to biased data
    - Reasons for missing information are not addressed
    - Missing at random assumption is rarely true

Estimate Values

- Missing at random not necessarily assumed
- More work to implement
- May Lead to less bias

- Downsides
  - Requires understanding of the relationships between variables and reasons for missing data
  - Choice of estimation method may affect the results

Extreme Values

Extreme Values

Definitions: The data points that are significantly different from the majority of data in a dataset. They may occur due to variability in the data or errors in data collections.

- **Univariate Outliers:** Extreme values in a single variable.
- **Multivariate Outliers:** Extreme values that occur in the context of multiple variables.

Purpose

- Impact on Analysis: Outliers can skew statistical results, affect mean and standard deviation, and distort data visualizations.

- Data Quality: Identifying outliers helps in detecting errors or unusual conditions.

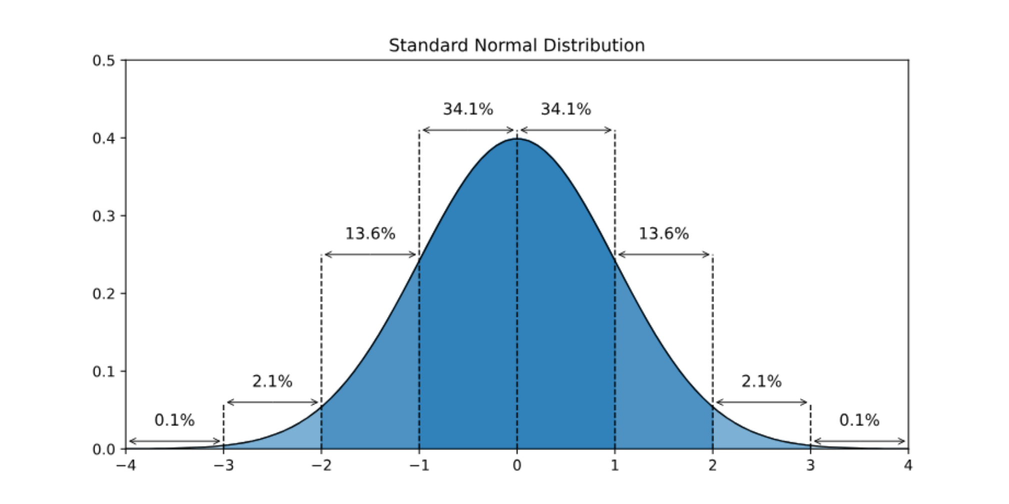
- Decision Making: Proper handling of outliers leads to more accurate models and better decision making.

Causes of Extreme Values

- **Natural Variability:** Some outliers occur naturally due to genuine variability in data (e.g., income, age).
- **Measurement Error:** Data entry errors, equipment malfunctions, or other issues can cause outliers.
- **Sampling Error:** Sometimes, outliers appear due to the way a sample was drawn.
- **Experimental Conditions:** Outliers can result from specific conditions in an experiment that were not controlled.

Statistical Methods

Z-Score: Explain how Z-scores can be used to detect outliers in normally distributed data. A Z-score above 3 or below -3 is often considered an outlier.



$$Z = \frac{x - \mu}{\sigma}$$

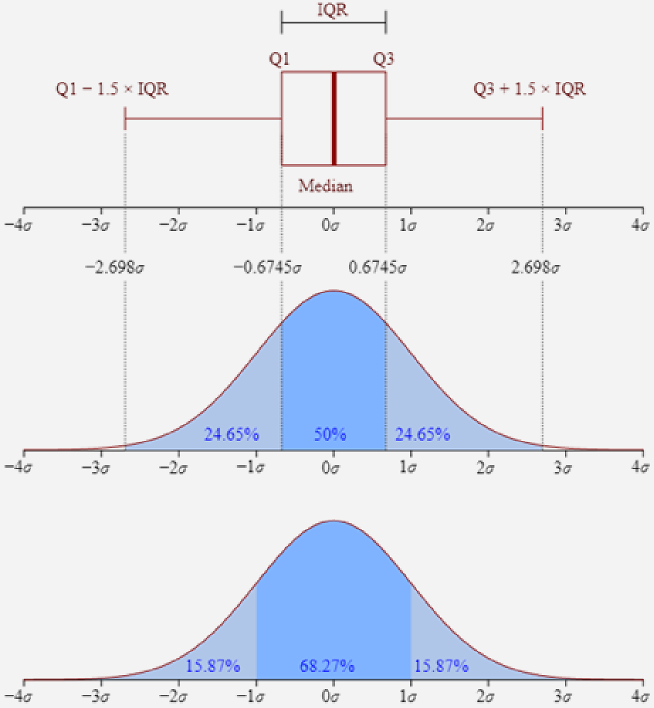
Statistical Methods

Interquartile Range(IQR): IQR is a measure of statistical dispersion, or spread, that describes the range within which the central **50%** of values in a dataset lie. It is the difference between the **third quartile (Q_3)** and the **first quartile (Q_1),** which represent the **25^{th}** and **75^{th}** percentiles, respectively.

$$IQR = Q_3 - Q_1$$

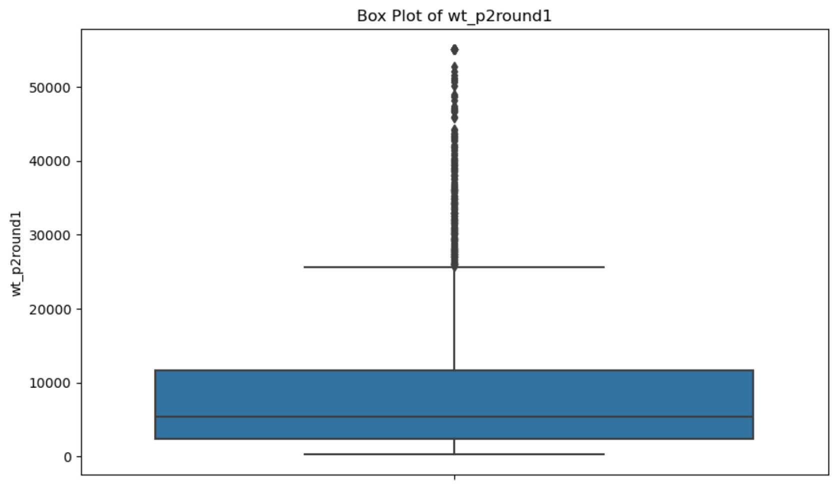
Non-Outlier Range: (Inner Fences)

[Q_1 - 1.5*IQR, Q_3 + 1.5*IQR]

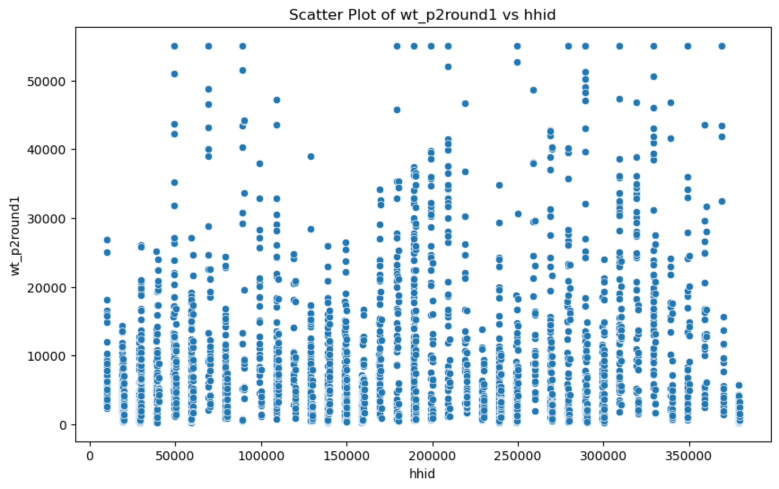


Visual Methods

Box Plot



Scatter Plot



Question

Does a data point being an outliter indicate an error?

Handling Outliers

- **Remove the Outlier:** If it’s an error, it might be reasonable to remove it.
- **Transform the Data:** Apply transformations (e.g., log, square root) to reduce the impact of outliers.
- **Keep and Report:** In some cases, keeping outliers might be important, especially if they carry valuable information (e.g., in fraud detection).

Key Concepts of Data Analytics

- Descriptive Analysis - What happened?
- Diagnostic Analysis - Why did it happen?
- Predictive Analysis - What will happen?
- Prescriptive Analysis - What should be done?

Descriptive Analysis

Purpose: Summarize historical data to understand what has happened.

Techniques:
Statistical analysis (mean, median, mode, standard deviation, etc.)
Data visualization
Data aggregation
Data mining

Examples:
Customer satisfaction surveys
Monthly sales reports
Website traffic summaries

Descriptive Analysis

- **Mean:** The mean (AKA the expected value) is the average of a set of numbers. It is calculated by summing all the numbers and then dividing by the count of numbers, and its mathematical notation is μ.

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- **Mode:** The mode is the value that appears most frequently in a data set. A set of numbers may have one mode, more than one mode, or no mode at all if no number repeats.

- **Median:** The median is the middle value in a set of numbers when they are arranged in ascending or descending order. If there is an even number of observations, the median is the average of the two middle numbers.

Descriptive Analysis

- **Standard Deviation:** The standard deviation measures the amount of variation or dispersion in a set of values. It indicates how much the values in the data set deviate from the mean on average.

- For a Population:

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}$$

- For a Sample:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2}$$

Diagnostic Analysis

Purpose: Investigate and explain why something happened.

Techniques: Hypothesis testing
Data discovery
Correlation analysis
Root cause analysis

Examples: Analyzing why a marketing campaign was successful or unsuccessful

Hypothesis Testing

Step 1: State the Hypothesis

- **Null Hypothesis (H_0):** This is the statement that there is no effect or no difference, and it represents the status quo. The goal is to test this hypothesis.

- **Alternative Hypothesis (H_1):** This is the statement that there is no effect or no difference, and it represents the status quo. The goal is to test this hypothesis.

Example:

$H_0: \mu = \mu_0$ (the population mean is equal to a specified value)

$H_1: \mu \neq \mu_0$ (the population mean is different from the specified value)

Hypothesis Testing

Step 2: Choose the Significance Level (α)

The significance level is the probability of rejecting the null hypothesis when it is actually true. Common choices for α are 0.05, 0.01, or 0.10.

- α=0.05 implies a 5% risk of concluding that a difference exists when there is no actual difference.

Step 3: Select the Appropriate Test

Choose the statistical test based on the type of data and the nature of the hypothesis.

- z-test: Used when the sample size is large ($n > 30$) or the population standard deviation is known.
- t-test: Used when the sample size is small ($n < 30$) and the population standard deviation is unknown.
- Chi-square test: Used for categorical data to assess how likely it is that an observed distribution is due to chance.
- ANOVA: Used when comparing means across three or more groups.

Hypothesis Testing

Step 4:Calculate the Test Statistic

Compute the test statistic (z, t, etc.) using the sample data.

Z-Statistics:

$$z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

T-Statistics (one-sample)

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Step 5: Determine the P-Value or Critical Value

P-Value: The probability of observing the test statistic or something more extreme, assuming the null hypothesis is true.

- Compare the p-value with α.
- If $p \leq \alpha$, reject the null hypothesis.
- If $p > \alpha$, fail to reject the null hypothesis.

Hypothesis Testing

Step 5: Determine the P-Value or Critical Value (Continue)

Critical Value: A threshold value that the test statistic is compared against to decide whether to reject the null hypothesis.

- Find the critical value(s) from statistical tables based on the chosen α level.
- If the test statistic is more extreme than the critical value, reject the null hypothesis.

Step 6: Make a Decision

- Reject H_0 : If the test statistic falls into the rejection region (or if the p-value is less than α).
- Fail to Reject H_0 : If the test statistic does not fall into the rejection region (or if the p-value is greater than α).

Hypothesis Testing

Step 7: State a Conclusion

- Write the conclusion in the context of the problem, clearly stating whether there is enough evidence to support the alternative hypothesis.
- Example:

"Based on the test, we reject the null hypothesis at the 0.05 significance level. There is sufficient evidence to conclude that the mean is different from μ_0 ."

Step 8: Consider the Practical Significance

- Even if a result is statistically significant, consider whether it is practically significant, meaning whether the effect size is large enough to be meaningful in a real-world context.

Predictive Analysis

Purpose: To predict future outcomes based on historical data and identify trends and patterns.

Techniques:
- Machine learning algorithms (regression, classification, clustering)
- Time series analysis
- Predictive modeling

Examples:
- Forecasting future sales
- Credit scoring

Regression Analysis (Linear Regression)

Definition: Linear regression is a statistical method for modelling the relationship between a dependent variable and one or more independent variables.

The Linear Regression Equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon$$

- y: Dependent variable (what we're trying to predict)
- $x_1, x_2, \ldots, x_n$: Independent variables (features)
- β_0: Intercept (the value of y when all $x$xs are 0)
- $\beta_1, \beta_2, \ldots, \beta_n$: Coefficients (weights of the features)
- ϵ: Error term (difference between the predicted and actual values)

Prescriptive Analysis

Purpose: To provide recommendations for actions that can be taken to affect desired outcomes. It could be short-term or long-term.

Techniques: Simulation modeling
Decision analysis
Optimization

Examples: Optimizing supply chain operations

Workshop Structure:

- Types of Data:
    - We'll start by identifying different types of data you will encounter. Knowing whether data is structured or unstructured will help you choose the appropriate analysis or processing method.

- Data Preparation:
    - We will use pandas and numpy in Python to clean and prepare our data. This stage is crucial for ensuring the accuracy of your analyses

Lets Get Started:

- Step 1 Setting the Environment
- Please open the Jupyter Notebook

More Practice:

- Step 2 Loading Data
- Please open the Jupyter Notebook

Data Cleaning Practice:

- Step 3 Cleaning Data
- Please open the Jupyter Notebook

Questions?

SWB is grateful for the partnership of University of Lagos Department of Statistics

https://StatisticsWithoutBorders.org

StatisticsWithoutBorders@gmail.com

UNIVERSITY OF LAGOS

IN DEED AND IN TRUTH

Σ STATISTICS WITHOUT BORDERS