**Week 1 Class 1: By Wangui Ngina, MSc, Lecturer and Data Scientist.**

**Time: Wednesday, September 27th, 2023 8:00 PM EAT.**

**Introduction to Data Science**

* + What is data science?
  + Why is data science important?
  + The different career paths in data science
  + The skills and knowledge required for a data science career

**Statistical, Mathematical, and Programming Concepts for Data Science**

* Statistics
  + Descriptive statistics
  + Inferential statistics
  + Probability
  + Linear algebra
  + Calculus
* Programming (Here discuss the major programming concepts in Python and SQL) that are used in Data Science) Example:
  + Python
  + SQL

**Examples: Python**

* Data structures. Data structures are used to store and organize data. Some common data structures used in data science include lists, dictionaries, and tuples.
* Algorithms. Algorithms are a set of instructions that are used to solve a problem. Some common algorithms used in data science include sorting algorithms, searching algorithms, and machine learning algorithms.
* Functions. Functions are blocks of code that can be reused. Functions are often used to perform repetitive tasks or to organize code.
* Object-oriented programming (OOP). OOP is a programming paradigm that allows programmers to create objects that have their own data and methods. OOP is often used in data science to create models and simulations.
* Data visualization. Data visualization is the process of representing data in a graphical way. Data visualization is often used in data science to communicate insights from data.

**Examples: SQL**

* SELECT statement. The SELECT statement is used to select data from a database table.
* WHERE clause. The WHERE clause is used to filter the results of a SELECT statement.
* JOINs. JOINs are used to combine data from two or more tables.
* GROUP BY clause. The GROUP BY clause is used to group the results of a SELECT statement.
* ORDER BY clause. The ORDER BY clause is used to sort the results of a SELECT statement.
* Subqueries. Subqueries are queries that are nested inside other queries.

======================================================================

**Statistics, Probability, Linear Algebra, Calculus, Python and SQL Programming Languages**

Note we are assuming all learners are beginners, so start with the introduction to data science.

**Please use real-world examples to illustrate the concepts.**

Example: A real-world examples of linear algebra in the field of data science:

**Linear Regression**:

* + **Scenario**: Imagine you want to predict the price of a house based on certain features like the size of the house (in square feet), the number of bedrooms, and the neighborhood's crime rate.
  + **Linear Algebra Use**: Linear regression involves finding a linear relationship between the input features and the target variable (house price in this case). Linear algebra is used to represent this relationship as a system of linear equations.
  + **Explanation**: Linear algebra helps create a model that estimates the house price by assigning weights to each feature (size, bedrooms, crime rate) and combining them linearly. This is achieved through matrix multiplication, where the feature values are multiplied by their respective weights and summed up. The result is an equation that predicts the house price.

**Principal Component Analysis (PCA)**:

* + **Scenario**: Consider a dataset with numerous features (e.g., measurements of a plant's height, width, and leaf size). You want to reduce the number of features while preserving as much of the original data's information as possible.
  + **Linear Algebra Use**: PCA is a dimensionality reduction technique that involves finding the principal components of the data using linear algebraic transformations.
  + **Explanation**: Linear algebra helps identify the most important directions (principal components) in the data. These components are linear combinations of the original features. By keeping only the most significant components, you reduce the dimensionality of the dataset while minimizing the loss of information. This is incredibly valuable for data visualization and simplifying complex datasets.

In all these examples, linear algebra provides the mathematical foundation to model relationships between variables, make predictions, and reduce the complexity of data. It is a fundamental tool in data science for solving real-world problems by representing and manipulating data in a way that facilitates analysis and decision-making.