Data Science using Python

Google Developer Student Clubs

Google Developer Student Clubs

UET Lahore New Campus

Speakers

Abdul Mughni

Google DSC

Data Science Lead

Hamza Atif

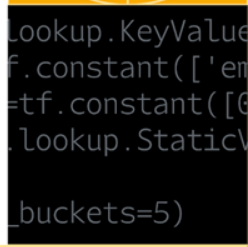
Google DSC

DS Co-Lead

Uzair Sarfraz

GDSC

Campus Lead



Organizers

M.Uzair Sarfraz

UET

GDSC Lead

Khubaib Ahsan Mustafa

Co-Lead

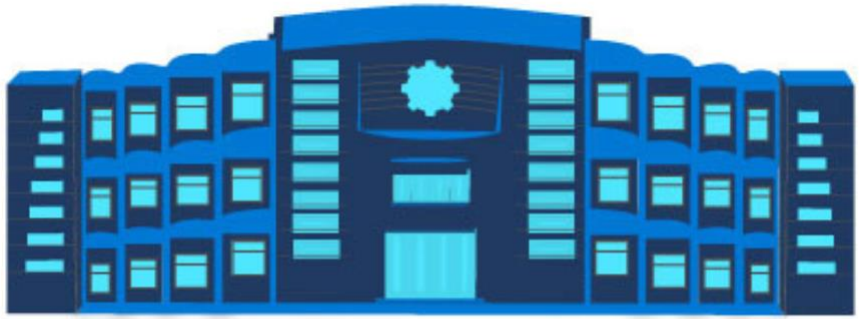
Muhammad Zubair Jameel Chughtai khan

Outreach Lead

Asfand Rasheed Khan

Management Lead

PARTNERS



MLSA UET LAHORE
NEW CAMPUS

Google Developer Student Clubs
University of Engineering and Technology lhr, New Campus

Python Libraries

Hamza Atif
Data Science Co-Lead

Agenda:

1. Introduction
2. Python Setup
3. Anaconda Environment Setup
4. Intro to Python Libraries
5. Exploratory Data Analysis with Python
6. QnA Session

Introduction to libraries

Python libraries are collections of pre-written code that provide a variety of functionalities and features.

What they are:

- Collections of modules, which are individual files containing Python code.
- Modules can contain functions, classes, and other types of objects.
- Libraries providing functionalities like data manipulation, web scraping, or machine learning.
- Libraries can also be domain-specific, focusing on specific areas like science, finance, or web development.

Overview:

Three essential Python libraries for data science:

- **Pandas:** A powerful library for data manipulation and analysis.
- **NumPy:** A library for efficient numerical computation.
- **Matplotlib:** A library for data visualization.

Pandas

- Open-source library for data manipulation and analysis.
- Provides powerful data structures like DataFrame.

Features of Pandas :

- Importing and exporting data from various formats (CSV, Excel, etc.)
- Data cleaning and manipulation.
- Descriptive statistics and aggregations.
- Data filtering and indexing.

NumPy

- Open-source library for efficient numerical computation.
- Provides high-performance n-dimensional arrays.

Features of NumPy :

- Mathematical operations on arrays
- Linear algebra routines
- Random number generation
- Fast data processing

Matplotlib

- Open-source library for data visualization.
- Creates various types of plots and charts.

Features of Matplotlib :

- Line plots, scatter plots, bar charts, histograms, etc.
- Customization of plot appearance (colors, labels, etc.)
- Integration with other libraries

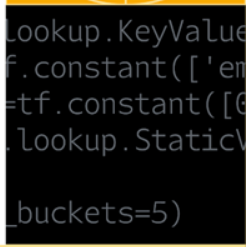
Google Developer Student Clubs
University of Engineering and Technology lhr, New Campus

EDA with Python

Abdul Mughni
Data Science Lead

Google Developer Student Clubs
University of Engineering and Technology Ihr, New Campus

Exploratory Data Analysis



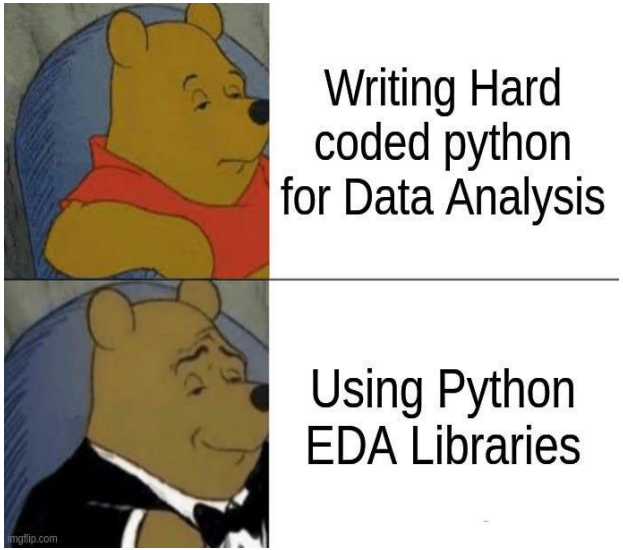
Exploratory Data Analysis

Exploratory Data Analysis (EDA) is an approach for data analysis that employs variety of techniques (mostly graphical) to

1. Maximize insight into a data set
2. Extract important variables
3. Detect outliers and anomalies
4. Develop models

EDA Goals

- The primary goal of EDA is to maximize the analyst's insight into a data set and into the underlying struture of a data set
- To get a "feel" for the data, the analyst also must know what is not in the data



Types of EDA

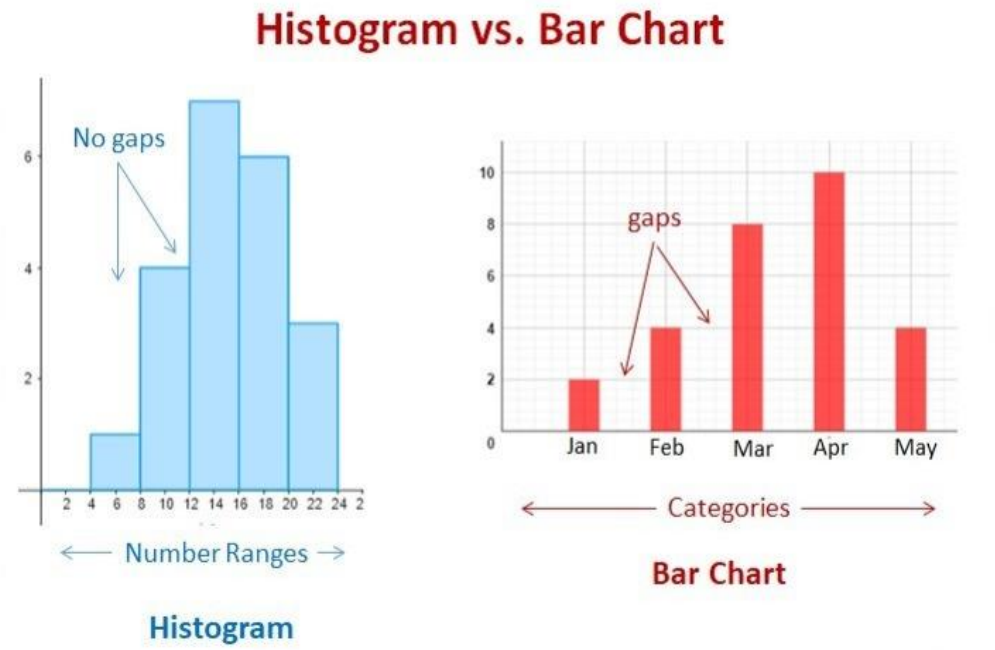
- Univariate Data Analysis
- Bivariate Data Analysis
- Multivariate Data Analysis

Univariate Data Analysis

Univariate data analysis are straightforward as we are dealing with only **one variable**.

The main purpose is to describe the data and find patterns that are present.

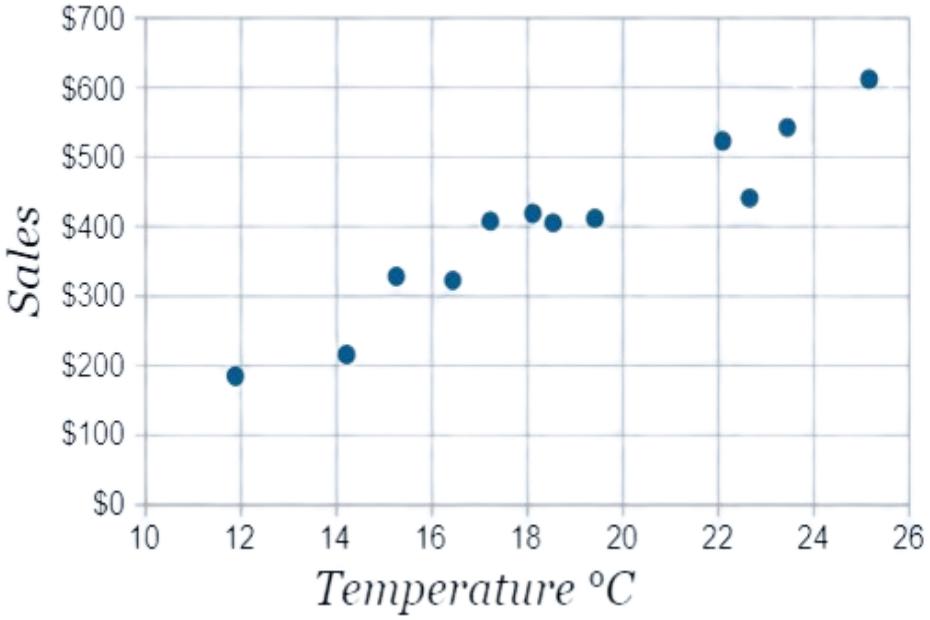
- Bar chart
- Pie chart
- Histogram



Bivariate Data Analysis

Bivariate data involves **two different variables** where we are concerned about investigating the causes and relationship between those 2 variables.

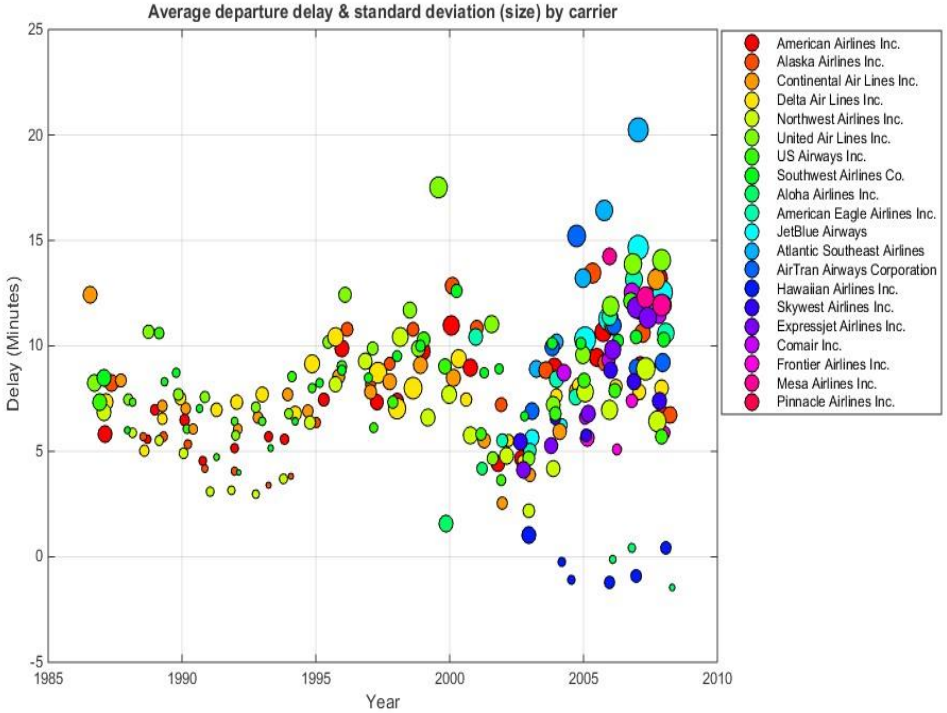
- Scatter Plot
- Line Plot
- Stacked Bar chart



Multivariate Data Analysis

Data which **involves 3 or more variables** are termed as Multivariate data. These are similar to bivariate but contains more than one dependent variable.

- Enhanced Basic Plots



Seaborn vs. Matplotlib

Google Developer Student Clubs

Seaborn vs. Matplotlib

Python's two most widely used data visualization libraries are Matplotlib and Seaborn. While both libraries are designed to create high-quality graphics and visualizations

- **Matplotlib** is a low-level plotting library that provides a wide range of tools for creating highly customizable visualizations.

- **Seaborn**, on the other hand, is a high-level interface for creating statistical graphics. It is built on top of Matplotlib and provides a simpler, more intuitive interface for creating common statistical plots

Google Developer Student Clubs