December 4th - 5th, 2024 | 9:00 a.m.
New Library Conference - UDSM

Explanatory Data Analysis Using Python

ABUBAKAR YUSSUF OMAR

Software Developer
EGAZ

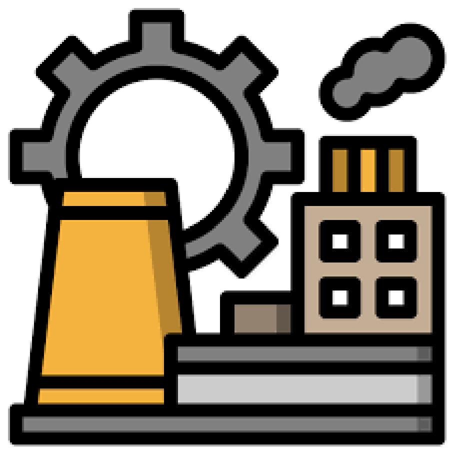
PYTHON CONFERENCE IN TANZANIA

Big Picture About Data Analysis

Data analytics involves looking at historical data and real time data to uncover trends, patterns and other descriptive information about an organization or a business.

It is used to present information in a way that is easy for non-technical decision makers to understand , It plays an important role in decision-making support for all types of organizations.

Advances in the collection, storage and retrieval of real-time data make possible the analysis of all types of data from many sources, such as documents, images, entries in social media or e-commerce sites, and even sensors inside home appliances..



Data Analysis answers the following questions:

“What happened?”. - **Descriptive analytics**

- Reporting general trends like revenue growth
- Reporting employee injuries

“What might happen in the future?” - **Predictive analytics**

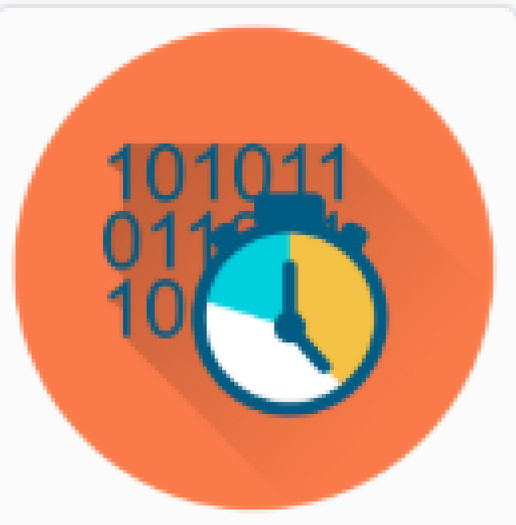
- Predicting customer preferences and recommending products to customers based on past purchases and search history

.

Data Analysis answers the following questions:

“ “What should be done next?.” - **Prescriptive analytics**

- Feld of GPS-based map and direction applications.
- Provide route options to a destination based on traffic volume, road conditions, and maximum speed.



“Why did this happen?”. - **Diagnostic analytics**

- Subscription cancellations, correlated with customer comments and ratings, to determine the most common reasons why users cancel subscriptions.

.

Dataset Repository

Data is now being collected and shared across many different organizations and in many different formats; a collection of data is referred to as a dataset.

It may be private or public dataset.

One of the most common formats used to package and exchange data is the Comma Separated Values (CSV) format. Often, datasets that are publicly available may be made up of multiple CSV files that contain related data. These CSV files can be imported into tools such as Excel for further investigation and analysis.

Bias in Data Analysis :

Bias is a natural tendency that all humans have, whether we are aware of it or not.

- **Confirmation bias**
- **Selection biases**
- **Interpretation bias**
- **Information bias**
- **Predictive bias**

Avoiding Bias:

- Be aware that bias exists.
- Validate your data sources and the methodology used to collect the data.
- Focus on larger patterns and trends
- Review your methods and data with others in your team
- Be open-minded and impartial in your analysis. Allow the data to inform your conclusions.

Example :

The Dataset concerns about scored by students from Math, Reading and writing subjects.

The aim is to understand influence of various factors on student performance.

Factors are:

- Economic.
- Personal.
- Social.

.

Student Performance on Exam Dataset

Requirements:

- Anaconda : https://www.anaconda.com/download/
- Numpy : https://pypi.org/project/numpy/
- Pandas: https://pypi.org/project/pandas/
- Matplotlib : https://pypi.org/project/matplotlib/
- Seaborn: https://pypi.org/project/seaborn/

Example :

The Dataset concerns about scored by students from Math, Reading and writing subjects.

The aim is to understand influence of various factors on student performance.

Factors are:

- Economic.
- Personal.
- Social.

.

THANK YOU FOR YOUR ATTENTIVE LISTENING.

.