UGANDA CHRISTIAN UNIVERSITY
A Centre of Excellence in the Heart of Africa

Data Science Lifecycle DSC8201

Week 1: Lecture 02 (MSDS_1:1)

Topic: *Data Management and Wrangling*

Dr. Daphne Nyachaki Bitalo
Department of Computing & Technology
Faculty of Engineering, Design & Technology

Breakdown of Course Outline

UGANDA CHRISTIAN UNIVERSITY
A Centre of Excellence in the Heart of Africa

Weeks	Course Goals	Focus Area	Topic Break Down	Practical Skills/ Tools
Week 1	Data Science Fundamentals & Ethics	Foundation & Project Setup	The Data Science Lifecycle (CRISP-DM, Cross-Industry Standard Process for Data Mining, or a similar framework), problem framing, project scoping, Data Privacy	Python, pandas, numpy setup, documentation
Week 1	Data Acquisition, Wrangling, and Storage	Data Engineering	Data ingestion from various sources (APIs, SQL, NoSQL, web scraping), data cleaning techniques (missing values, outliers), feature engineering basics,	ETL/ELT pipeline design, Pandas for data manipulation, Basic use of an object storage system (e.g., S3).
Week 2	Exploratory Data Analysis & Visualization	Analysis & Storytelling	Hypothesis generation and testing, descriptive statistics, data visualization principles, statistical inference (t-tests, ANOVA), communicating data stories.	Matplotlib, Seaborn, Plotly for visualization, statistical software (e.g., StatsModels, SciPy), Storytelling with data

Lecture Objectives and Learning outcomes

The Objectives of this lecture are to learn:

- Data management
- Data pre-processing
- Hand's on ETL pipeline flow

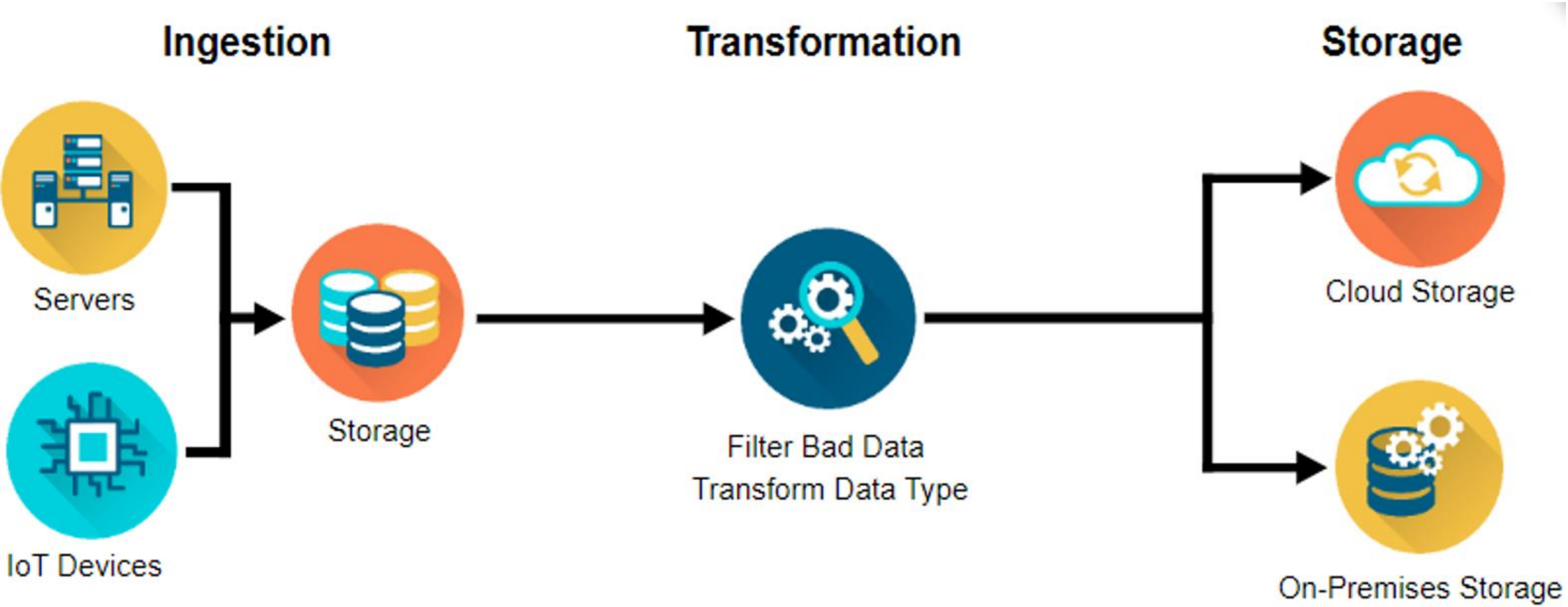
By the end of this lecture week, students should be able to:

- **Implement and manage the entire Data Science Lifecycle through ETL pipeline design**

Lecture Overview

- Data Management
- Data pre-processing
- Exploratory Data Analysis
- Describing Data (ETL pipeline)
- Object Storage systems

Understanding data management



Ingestion (Extraction)

Data engineers ingest two primary sources of data;

1. Batches of data from servers or databases (**batch** ingestion).

An example of batch ingestion is a game company that wants to examine the relationship between subscription renewals and customer support tickets. It could ingest all the related data on a daily or weekly basis. It doesn't need to access and analyze data immediately after a support ticket is closed or a subscription is renewed.

Ingestion (Extraction)

2. Real-time events happening in the world and streaming from the world of devices (**streaming** ingestion).

An example of streaming ingestion is when you request a ride from a ride share service. The company combines streams of data (e.g. historical data, real-time traffic data, and location tracking) to make sure you get a ride from the driver who is closest to you at the time.

Transformation

There are two main issues to deal with here;

1. Data often needs to be cleaned up.

Missing values, dates can be in the wrong format and data quickly gets outdated.

You might have gathered data on individuals who have changed roles or companies. So, all this data needs to be updated.

There might be data outliers that need to be handled as well

Transformation

2. Transforming the data so that its structure aligns with the system needed to allow accurate analyses.

For example;

You might want to figure out your company's best selling products every month. But the data may only contain each product's sale date. You would need to transform the data by creating a number of sales per month variable or total monthly sales.

Loading

After transforming data, it needs to be stored in places and forms, making it easy for analysts to run reports on weekly sales and data scientists to deduce insights and create predictive recommendation models.

Data security, or managing data access so that people who should be accessing the data can efficiently, and keeping out people who shouldn't.

There are two **primary locations** for businesses to store their data;

1. On-premises

2. In the cloud but often, companies use a hybrid of both.

The term "on-premises" refers to hardware on an organization's servers and infrastructure - usually physically on site.

Data Pre-processing

Data preprocessing is a crucial step in the data science pipeline, involving various techniques to prepare raw data for analysis. It ensures data quality, consistency, and suitability for modeling.

It involves Cleaning, Integration, Transformation, and Reduction

Data pre-processing

Cleaning

- **Handling missing values:** Imputation techniques (mean, median, mode, regression, etc.) or deletion.
- **Dealing with outliers:** Identification (statistical methods, visualization) and removal or correction.
- **Noise reduction:** Smoothing techniques (e.g., moving average) to remove noise or inconsistencies.
- **Data correction:** Identifying and correcting errors or inconsistencies in the data.

Data pre-processing

Integration

- **Merging datasets:** Combining multiple datasets based on common keys or identifiers.
- **Data standardization:** Ensuring consistency in data formats, units, and encoding.
- **Entity resolution:** Identifying and merging duplicate records representing the same entity.

Key for descriptive and predictive analysis

Data pre-processing

Transformation

- **Normalization:** Scaling data to a specific range (e.g., 0-1) to improve model performance (e.g. converting categorical data).
- **Feature engineering:** Creating new features from existing ones to capture relevant information.
- **Aggregation:** Combining multiple data points into a single value (e.g., calculating averages or sums).
- **Discretization:** Converting continuous data into discrete categories.

Data pre-processing

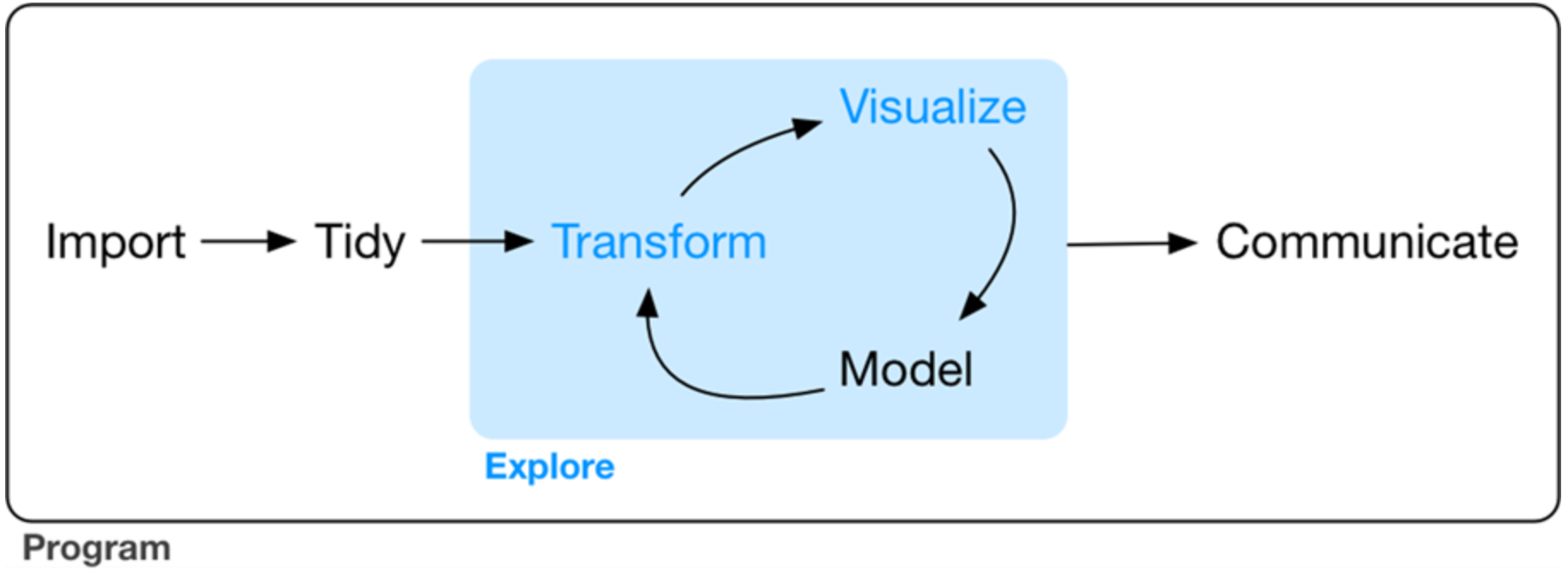
Reduction

- **Dimensionality reduction:** Reducing the number of features while preserving essential information.
- **Feature selection:** Choosing the most relevant features for analysis.
- **Sampling:** Selecting a subset of data for analysis to reduce processing time.

Key for predictive analysis

Steps of data exploration

Data exploration is the art of looking at your data, rapidly generating hypotheses, quickly testing them, then repeating again and again and again.



Exploratory Data Analysis (EDA)

An interactive cycle;

1. Generates questions about your data.
2. Searches for answers by visualizing, transforming, and modeling your data.
3. Uses what you learn to refine your questions and/or generate new questions.

There is no rule about which questions you should ask to guide your research. However, two types of questions will always be useful for making discoveries within your data. You can loosely word these questions as:

1. **What type of variation occurs within my variables? (i.e. variability)**
2. **What type of covariation occurs between my variables? (i.e. central tendency)**

All of the above questions generate **DESCRIPTIVE** and NOT predictive analyses

Central tendency

Central tendency is a measure that best summarizes the data and is a measure that is related to the center of the data set.

Described by the statistics Mode, Median and Mean

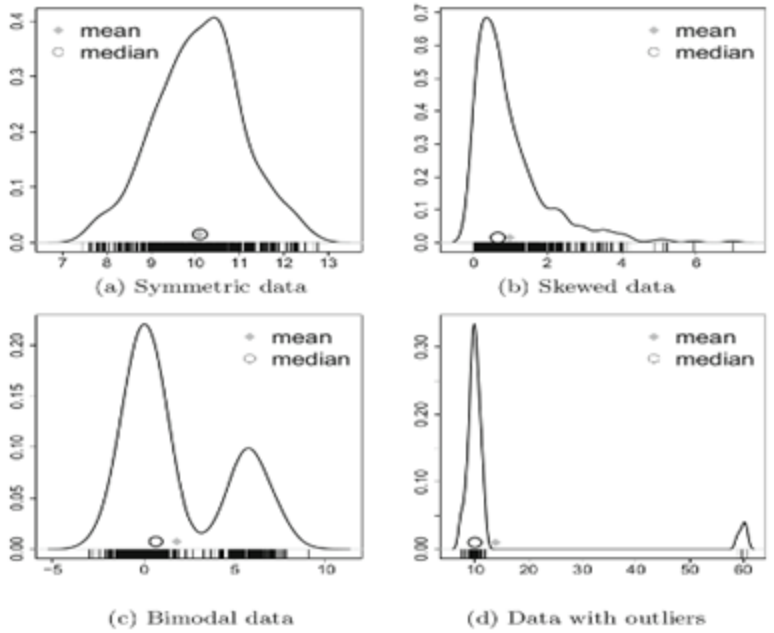


Fig. 3.1 Arithmetic mean and median for different data

Variability

Measures of variability are the measures of the spread/dispersion of the data.

Described by range, interquartile range, variance, standard deviation, and more

Variance is one of the most important measures in statistics

Covariation- two or more variables vary in a related manner. The best way to discover covariation is to visualize the relation.

The variance of a population is

$$\sigma^2 = \frac{\Sigma(X-u)^2}{N}$$

The variance of a sample is

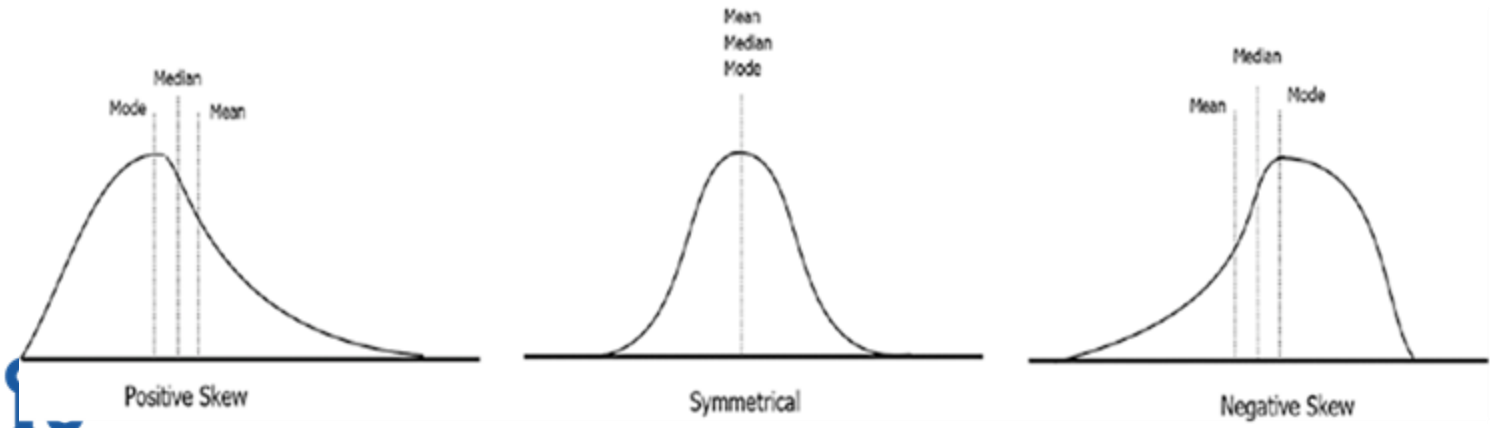
$$S^2 = \frac{\Sigma(X-\bar{X})^2}{n-1}$$

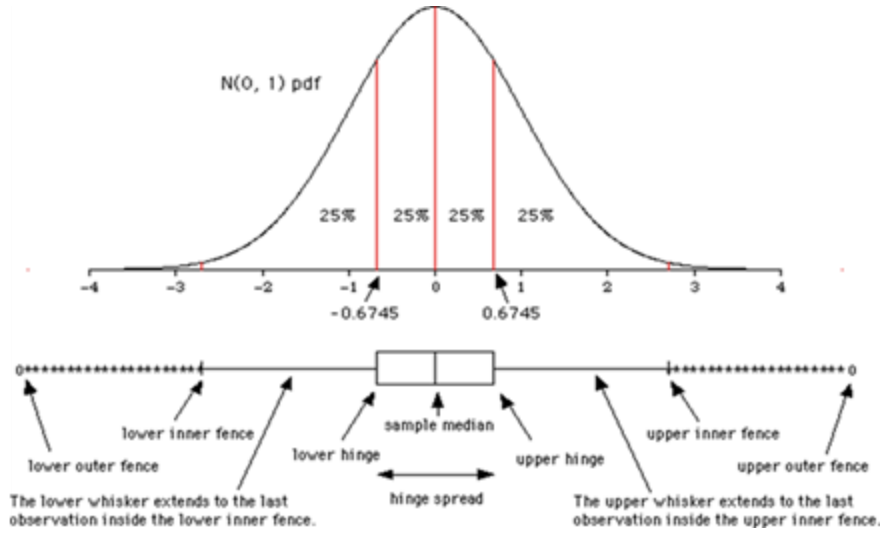
Defining the descriptive statistics

1. **Measure of central tendency: mean, median, mode**

Measures the "average" or the "middle" of your data. The most commonly used measures include:

- the mean: the average value. It's sensitive to outliers.
- the median: the middle value. It's a robust alternative to mean.
- and the mode: the most frequent value





P.O. Box 4, Mukono, Uganda, Plot 67-173, Bishop Tucker Road, Mukono Hill | Tel: +256 (0) 312 350 800 Email: info@ucu.ac.ug Web: https://ucu.ac.ug
Founded by the Province of the Church of Uganda. Chartered by the Government of Uganda

Defining the descriptive statistics

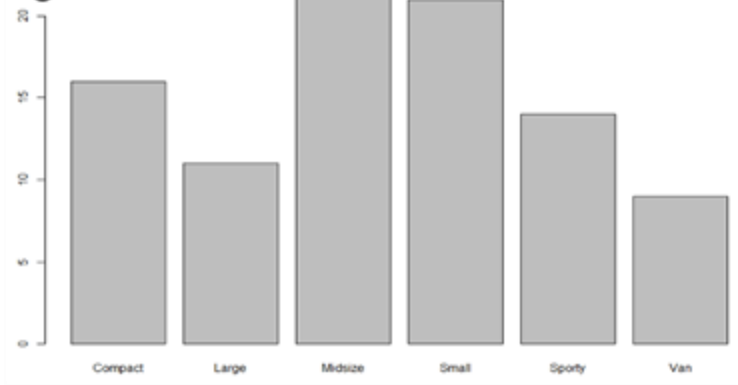
2. Measure of variability

Measures of variability gives how "spread out" the data are.

- Range: minimum & maximum
- Range corresponds to biggest value minus the smallest value. It gives you the full spread of the data.

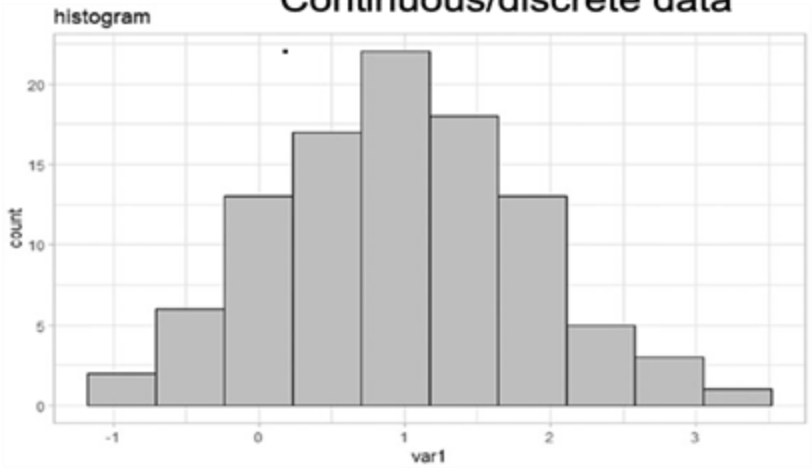
Bar Chart/ Bar plot

Categorical data



Histogram/ Density plot

Continuous/discrete data

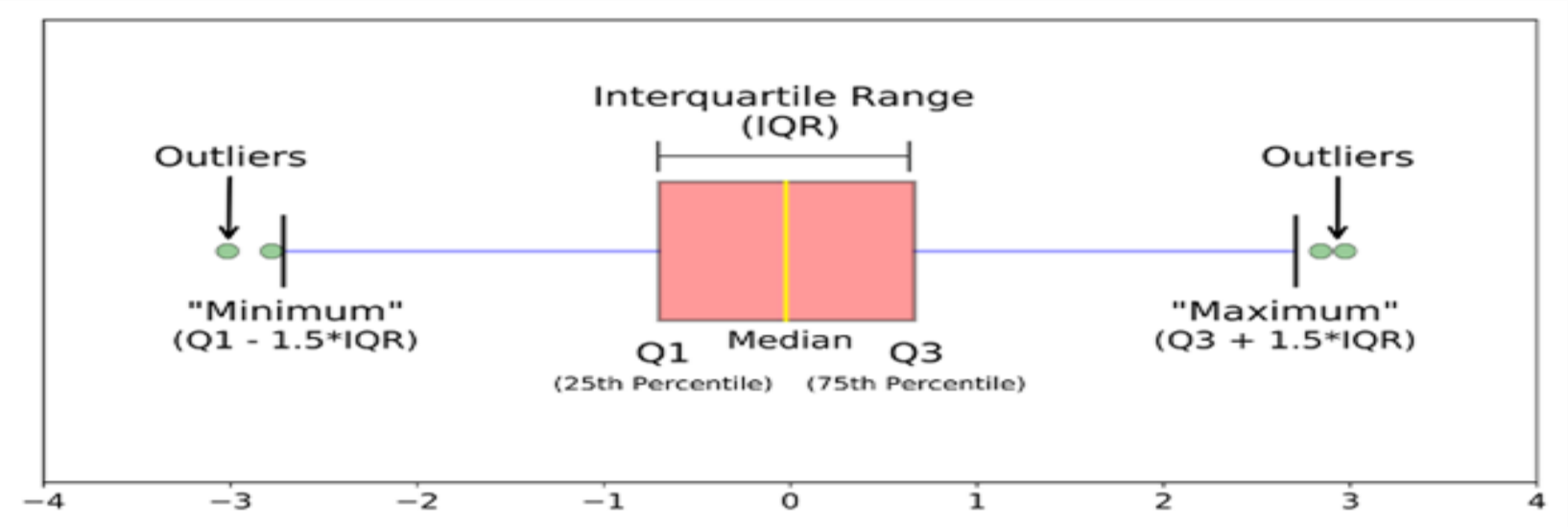


P.O. Box 4, Mukono, Uganda, Plot 67-173, Bishop Tucker Road, Mukono Hill | Tel: +256 (0) 312 350 800 Email: info@ucu.ac.ug Web: https://ucu.ac.ug
Founded by the Province of the Church of Uganda. Chartered by the Government of Uganda

Defining the descriptive statistics

3. Interquartile Range (IQR)

Quartiles divide the data into 4 parts. IQR corresponds to the difference between the first and third quartiles - is sometimes used as a robust alternative to the standard deviation.



P.O. Box 4, Mukono, Uganda, Plot 67-173, Bishop Tucker Road, Mukono Hill | Tel: +256 (0) 312 350 800 Email: info@ucu.ac.ug Web: https://ucu.ac.ug
Founded by the Province of the Church of Uganda. Chartered by the Government of Uganda

Purpose of each descriptive statistic

1. Range. It's not often used because it's very sensitive to outliers.
2. IQR. It's pretty robust to outliers.
3. Variance. It's completely uninterpretable because it doesn't use the same units as the data. It's almost never used
4. Standard deviation. This is the square root of the variance. It's expressed in the same units as the data. The standard deviation is often used in the situation where the mean is the measure of central tendency.
5. Median. It's a robust way to impute for missing data, for data with outliers. Used with continuous data that is not normally distributed.
6. Mode. Imputes for missing data that is categorical
7. Mean. Imputes for missing continuous data that is normally distributed

Testing for normal distribution

Rule of thumb:

Many of the statistical tests including correlation, regression, t-test, and analysis of variance (ANOVA) assume some certain characteristics about the data. They require the data to follow a normal distribution or Gaussian distribution. These tests are called parametric tests, because their validity depends on the distribution of the data.

Testing for normal distribution

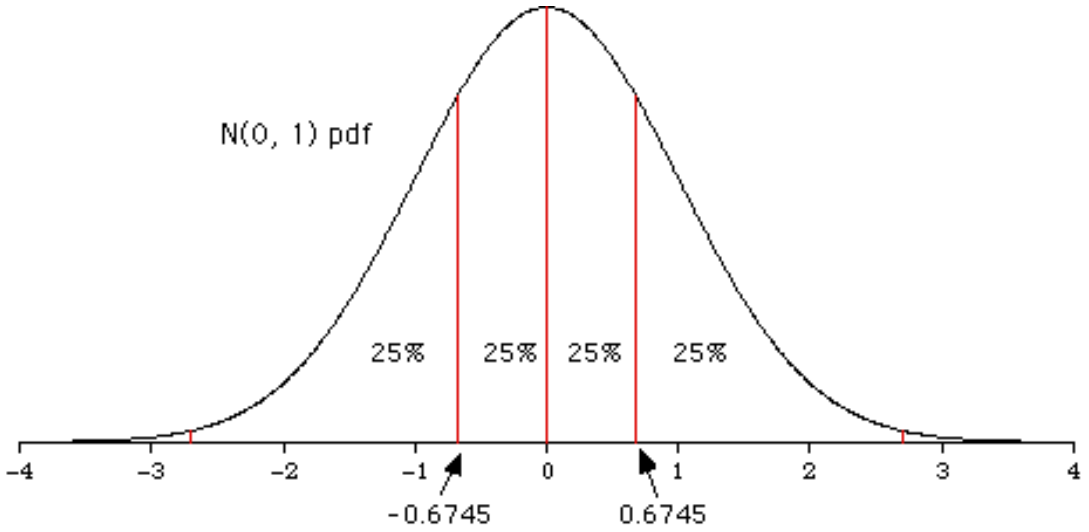
So the distribution of data has to be testing using the following:

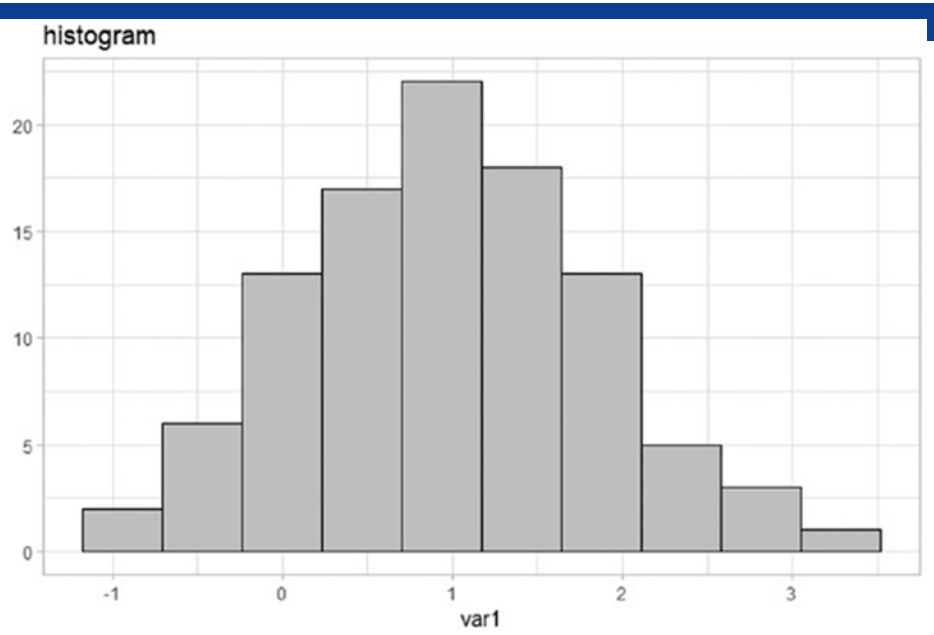
1. Visual inspection [normal plots (histogram), Q-Q plot (quantile-quantile plot)], boxplots)
2. Significance tests. Such as Shapiro-Wilk's test (samples >3< 5000) and Kolmogorov-Smirnov (K-S) test (samples > 5000)

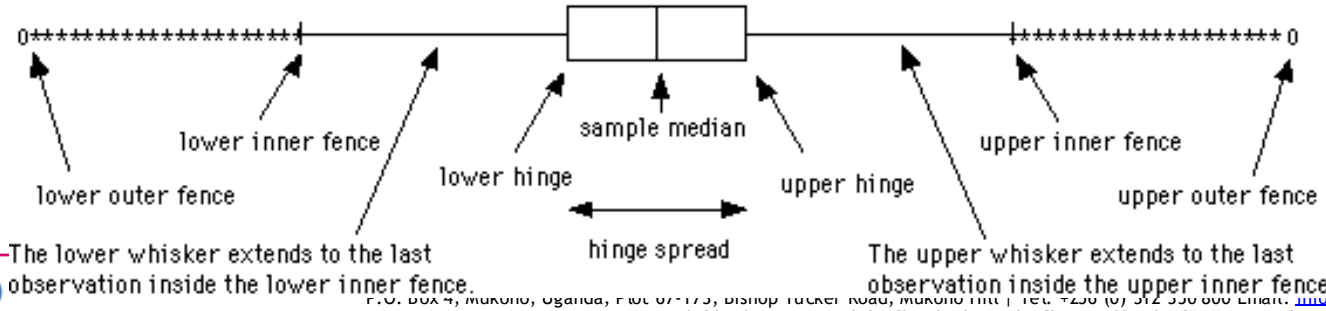
Data distribution also informs how missing data is handled

Central tendencies/normal distribution

The normal distribution = Gaussian distribution
Using visuals







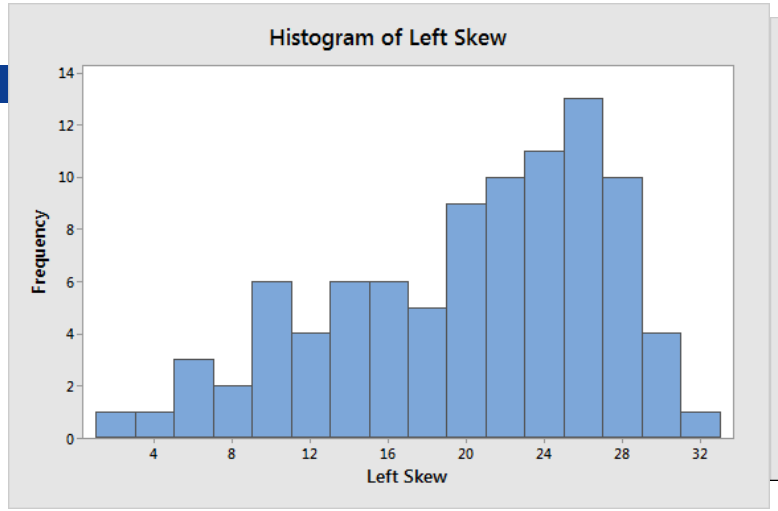
Negatively/Left skewed

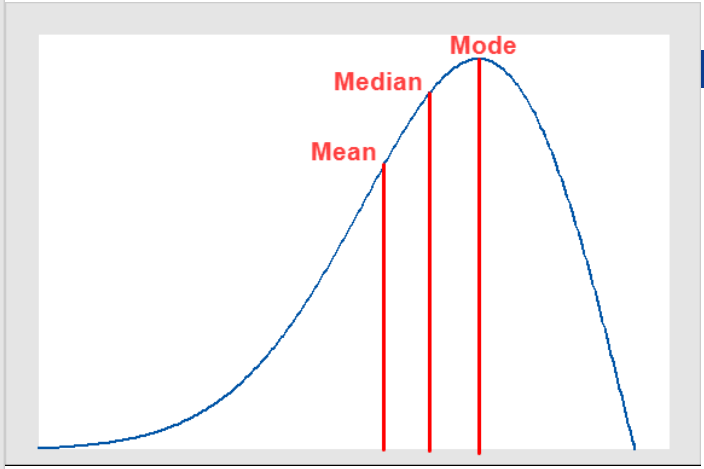
Left skewed distributions occur when the long tail is on the left side of the distribution

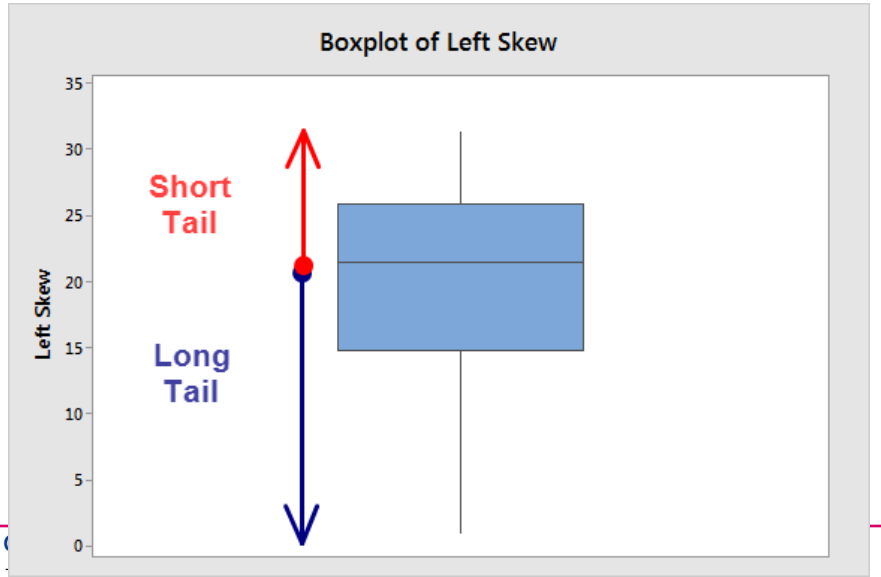
A box plot also displays this

As opposed to having the "bell-curve" of normal distribution;

- The mean is less than the median
- The mean underestimates the most common values in a negatively skewed distribution







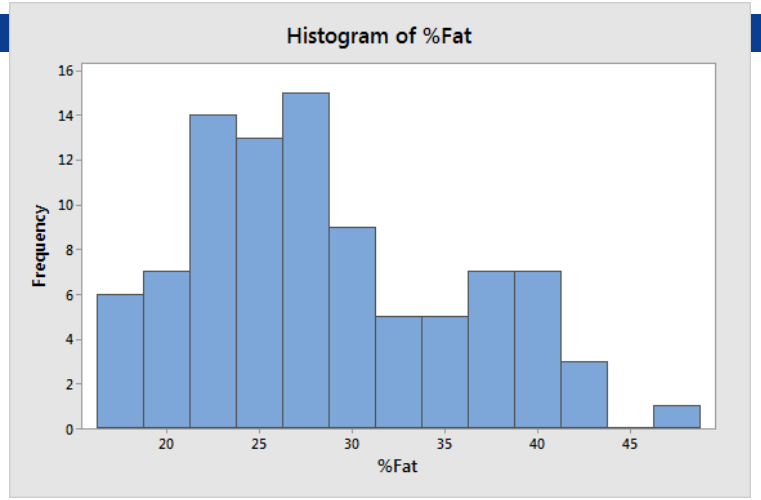
Positively/Right skewed

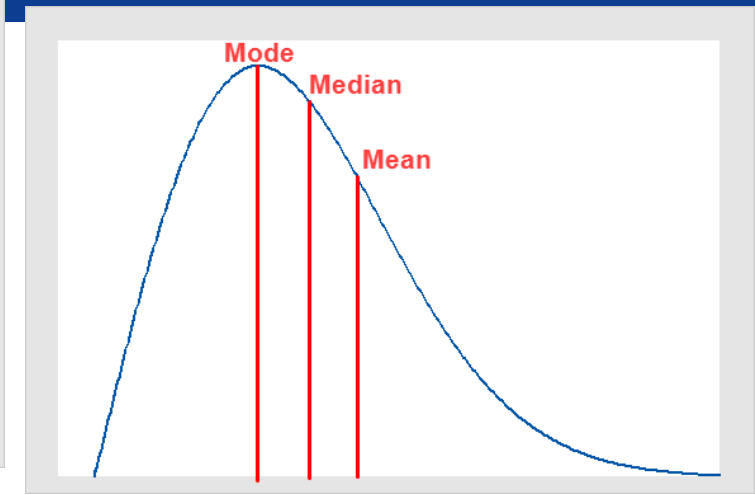
Right skewed distributions occur when the long tail is on the right side of the distribution

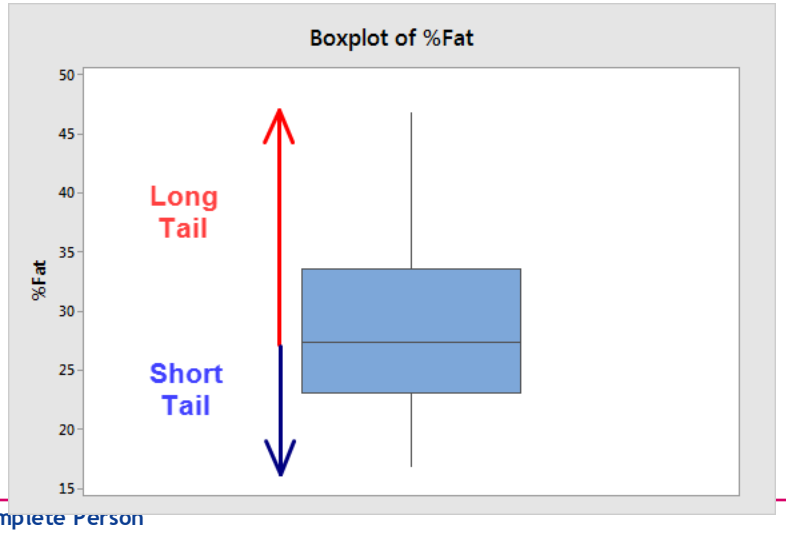
A box plot also displays this

As opposed to having the "bell-curve" of normal distribution;

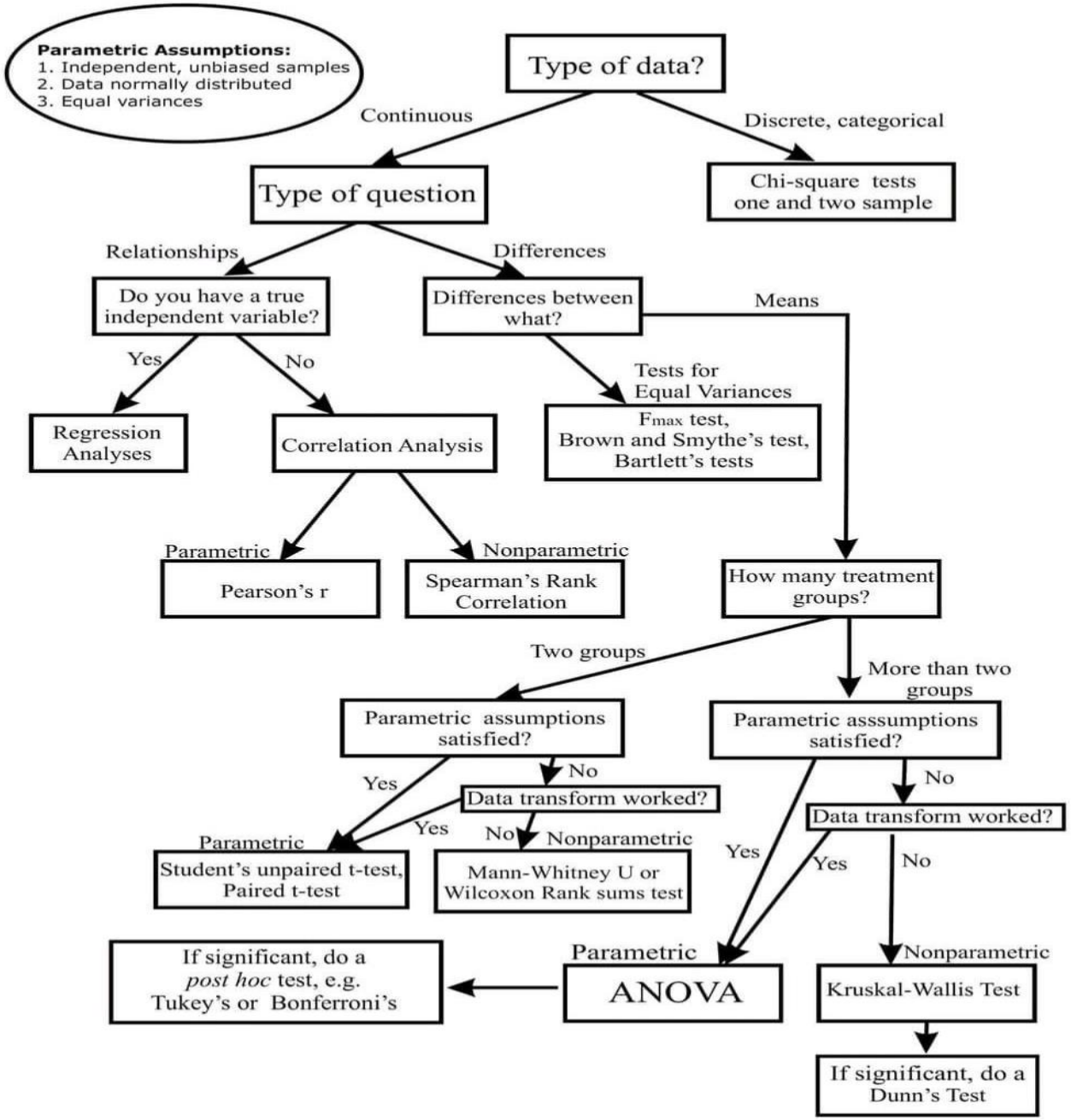
- The mean is greater than the median
- The mean overestimates the most common values in a positively skewed distribution







Flow Chart for Selecting Commonly Used Statistical Tests



Exercises using python

Practice with the Data_Preprocessing Notebook

Practice with the Exploratory Data Analysis Notebook

Group Exercise using python

1. Import the provided datasets in to visual studio ("Cassava_Yield_Data.xlsx" and "Bike_Sales.xlsx")
2. Explore the datasets by identifying the data type of each variable.
3. Which variables are observations and which are samples?
4. Which variables are populations? (This will be guided by the questions you pose)
5. Transform both datasets to handle missing data.
6. Save the transformed datasets as .csv files with your last name.
7. Using the above saved datasets, describe the central tendency of two continuous variables.
8. Generate graphs to show the distribution of one continuous variable and one categorical variable.

Object Storage Systems

UGANDA CHRISTIAN UNIVERSITY
A Centre of Excellence in the Heart of Africa

OSS's are ideal for data science due to their high scalability, ability to store vast amounts of unstructured data, and rich metadata capabilities.

- **Vast storage formats:** Support structured and unstructured data.
- **Accessibility:** Via standard APIs, python libraries.
- Examples: Amazon S3, Google Cloud, Azure Blob etc.

Key for data storage and management

OSS: Practical using Boto3

Boto3 is a python library that interacts with the Amazon S3 API.

- **Prerequisites:** Credentials to S3 system,
- Bash and boto3 library.

Which library would you use to access google drive?

UGANDA CHRISTIAN UNIVERSITY
A Centre of Excellence in the Heart of Africa

Uganda Christian University
P.O. Box 4 Mukono, Uganda
Tel: 256-312-350800
https://ucu.ac.ug/ Email: info@ucu.ac.ug.
@ugandachristianuniversity @UCUniversity
@UgandaChristianUniversity

Department of Computing & Technology
FACULTY OF ENGINEERING, DESIGN AND TECHNOLOGY
Tel: +256 (0) 312 350 863 | WhatsApp: +256 (0) 708 114 300
@ucucomputeng @ucu_ComputEng
https://cse.ucu.ac.ug/ Email: dct-info@ucu.ac.ug