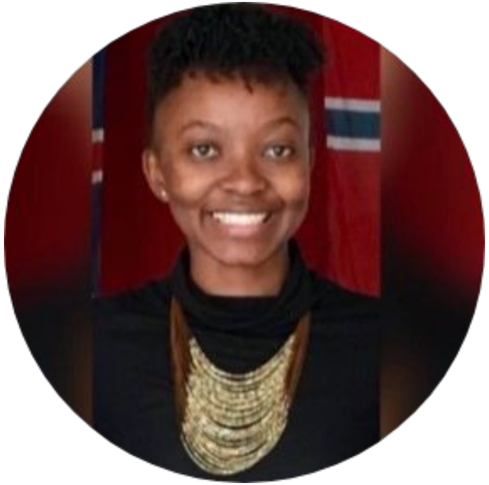
DSA Kimberley 2021

Women and Data in Africa

Data Analytics in Python

October 4, 2021

About us...



Dina Machuve, PhD

Senior Lecturer, Nelson Mandela African Institution of Science and Technology

Neema Mduma, PhD

Lecturer, Nelson Mandela African Institution of Science and Technology

Agenda

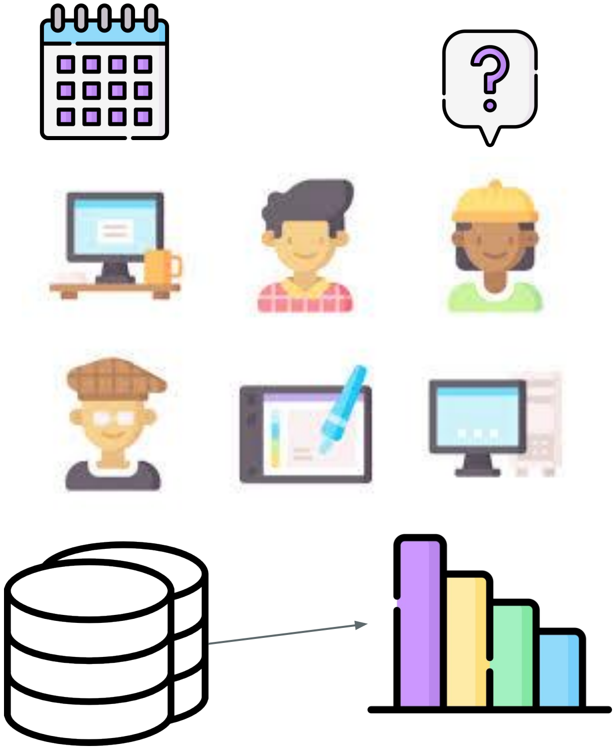
1. What is Data Analytics?
   a. Data Analytics Pipeline
   b. Python libraries for Data Analytics

2. How we can use Data Analytics to find insights on Financial Inclusion in Africa?
   a. Define the research question
   b. Data Validation and Cleaning
   c. Exploratory Analysis

3. Coding example on Data Analytics for Financial Inclusion in Africa

What is Data Analytics?

What is Data Analytics?



Data Analytics is the process of exploring and analyzing large datasets to make predictions and data-driven decision making.

Source: https://www.simplilearn.com/tutorials/data-analytics-tutorial/data-analytics-with-python

Data Science vs Data Analytics

Data Science

uses scientific methods and algorithms to extract knowledge and insights from structured and unstructured data.

Data Analytics

the act of inspecting datasets to infer conclusions from the information using specialized systems and software. It focuses on specific areas with specific goals.

Data Science and Data Analytics Overlap

	Data Analytics	Data Science
Machine Learning & AI	x	√
Statistics	√	√
Visualization	√	√
Data Wrangling & Mining	√	√
Reporting	√	x

Data Analytics Pipeline

Data Analytics Pipeline

~80% of your time as a data scientist is spent here, preparing the data for analysis

Machine learning takes place during the modeling phase.

Research Question

A research question is the question we want our model to answer.

Examples of research questions:

- Does this patient have malaria?
- Can we monitor illegal deforestation by detecting chainsaw noises in audio streamed from rainforests?
- Does this chicken have Newcastle disease?

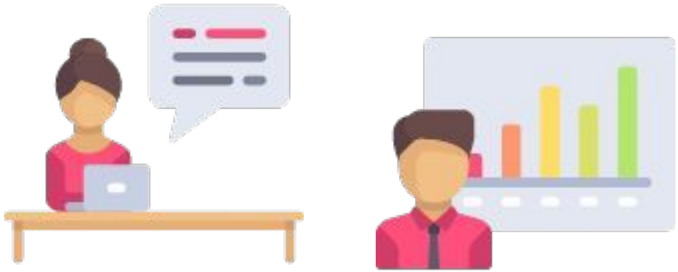
We may have a question in mind before we look at the data, but we will often use our exploration of the data to develop or refine our research question.

What comes first, the chicken or the egg?

Data Validation & Cleaning

Data Cleaning

Why do we need to validate and clean our data?



Data often comes from multiple sources

- Do data align across different sources?

Data is created by humans

- Does the data need to be transformed?
- Is it free from human bias and errors?

Data Cleaning

Data cleaning involves identifying any issues with our data and confirming our qualitative understanding of the data.

Missing Data
Is there missing data?
Is it missing systematically?

Data Type
Are all variables the right type?
Is a date treated like a date?

Times Series Validation
Is the data for the correct time range?

Data Range
Are all values in the expected range?

Data Cleaning

Missing Data

Is there missing data? Is data missing at random or systematically?

Very few datasets have no missing data; most of the time you will have to deal with missing data.

The first question you have to ask is what type of missing data you have.

Missing completely at random: no pattern in the missing data. This is the best type of missing you can hope for.

Missing at random: there is a pattern in your missing data but not in your variables of interest.

Missing not at random: there is a pattern in the missing data that systematically affects your primary variables.

Data Cleaning

Missing Data

Is there missing data? Is data missing at random or systematically?

Example: You have survey data from a random sample from high school students in Kenya. Some students didn't participate:

Some students were sick the day of the day of the survey

Some students declined to participate, since the survey asks about grades

If data is <u>missing at random,</u> we can use the rest of the nonmissing data without worrying about bias!

If data is missing in a non-random or <u>systematic</u> way, your nonmissing data may <u>be biased</u>

Data Cleaning

Missing Data

Is there missing data? Is data missing at random or systematically?

Example: You have survey data from a random sample from high school students in Kenya. Some students didn't participate:

Some students were sick the day of the day of the survey

If data is missing at random, we can use the rest of the nonmissing data without worrying about bias!

Some students declined to participate, since the survey asks about grades

If data is missing in a non-random or systematic way, your nonmissing data may be biased

Data Cleaning

Missing Data

Sometimes you can **replace** missing data

- Drop missing observations.
- Populate missing values with average of available data
- Impute data

What you should do depends heavily on what makes sense for your research question, and your data.

Data Cleaning

Missing Data

Common Imputation Techniques

Use the average of nonmissing values	Take the average of observations you do have to populate missing observations - i.e., assume that this observation is also represented by the population average
Use an educated guess	It sounds arbitrary and often isn't preferred, but you can infer a missing value. For related questions, for example, like those often presented in a matrix, if the participant responds with all "4s", assume that the missing value is a 4.
Use common point imputation	For a rating scale, using the middle point or most commonly chosen value. For example, on a five-point scale, substitute a 3, the midpoint, or a 4, the most common value (in many cases). This is a bit more structured than guessing, but it's still among the more risky options. Use caution unless you have good reason and data to support using the substitute value.

Source: Handling Missing Data

Exploratory Analysis

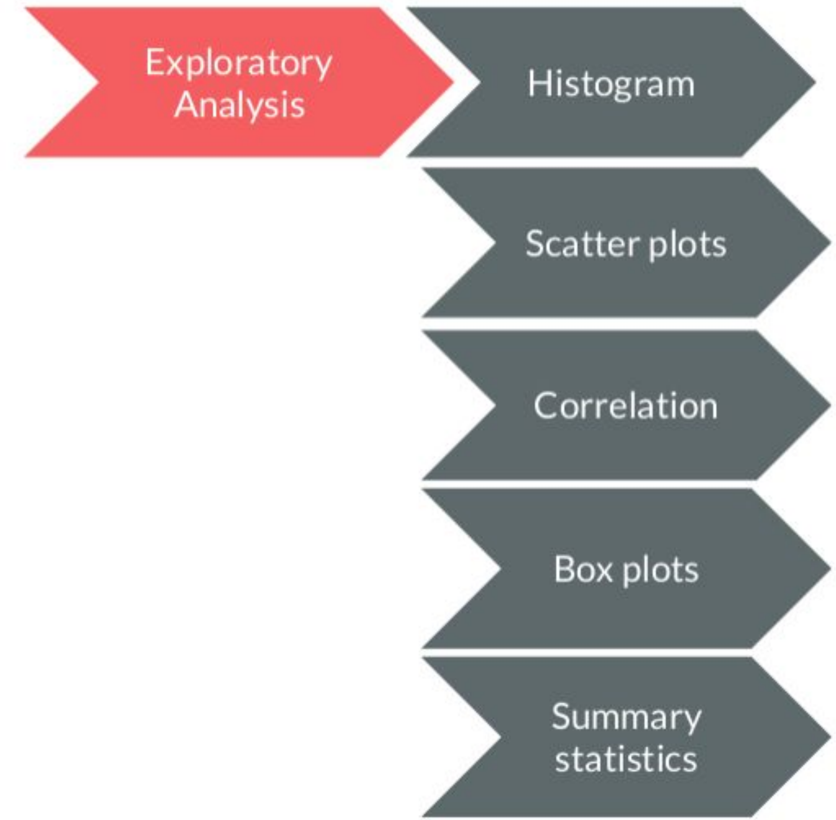
Exploratory Analysis

The goal of exploratory analysis is to better understand your data.

Exploratory analysis can reveal data limitations, what features are important, and inform what methods you use in answering your research question.

This is an indispensable first step in any data analysis!

Let's explore our data!



Once we have done some initial validation, we explore the data to see what models are suitable and what patterns we can identify.

The process varies depending on the data, your style, and time constraints, but typically exploration includes:

- Histogram
- Scatter plots
- Correlation tables
- Box plots
- Summary statistics
- Mean, median, frequency

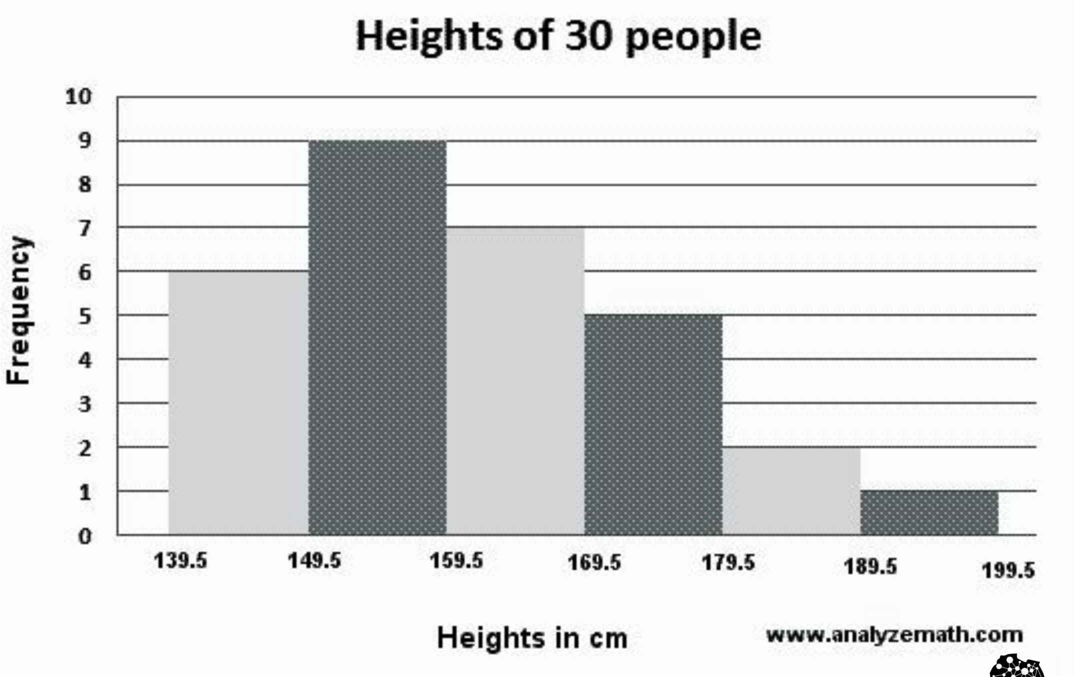
Exploratory Analysis

Histogram

Histograms tell us about the distribution of the feature.

A histogram shows the **frequency distribution** of a continuous feature.

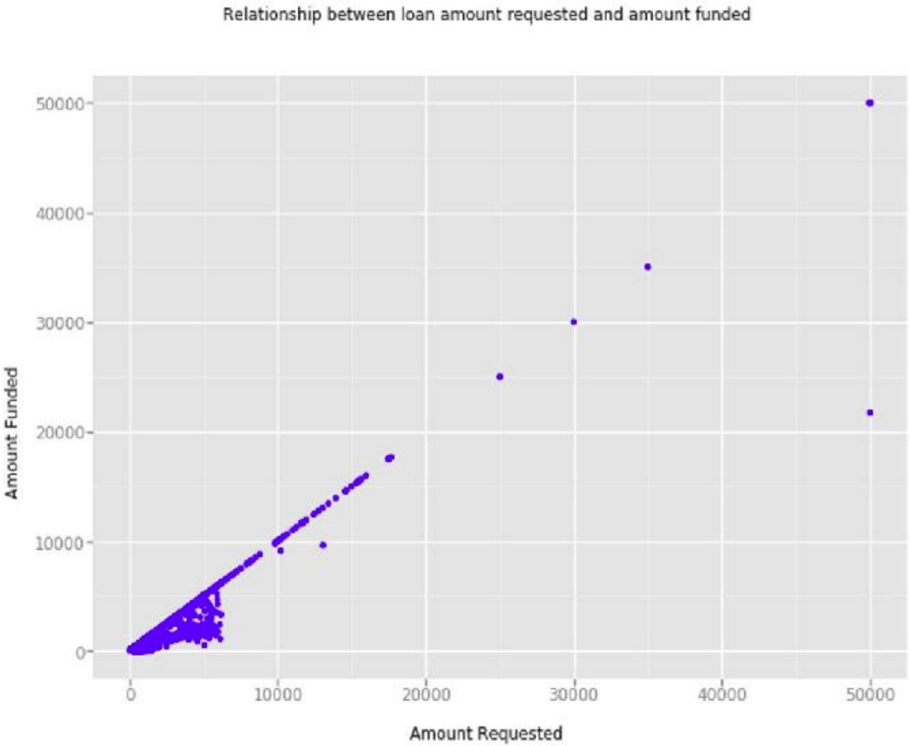
Here, we have height data of a group of people. We see that most of the people in the group are between 149 and 159 cm tall.



Exploratory Analysis

Scatter Plots

Scatter plots provide insight about the relationship between two features.

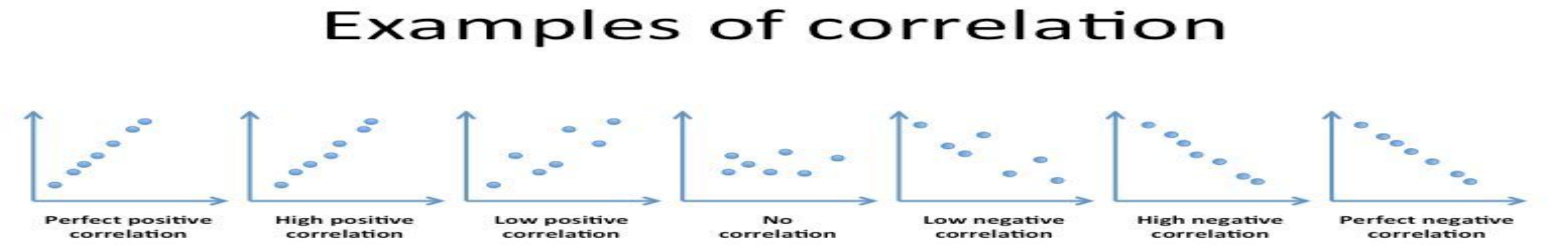


Scatter plots visualize relationships between any two features as points on a graph. They are a useful first step to exploring a research question.

Here, we can already see a positive relationship between amount funded and amount requested.
What can we conclude?

Exploratory Analysis > Correlation

Correlation is a useful measure of the strength of a relationship between two variables. It ranges from -1.00 to 1.00



1.00 0.88 0.60 0.00 -0.55 -0.78 -1.00

Exploratory Analysis

Correlation

Correlation does not equal causation

Let's say you are an executive at a company. You've gathered the following data:

X = $ spent on advertising

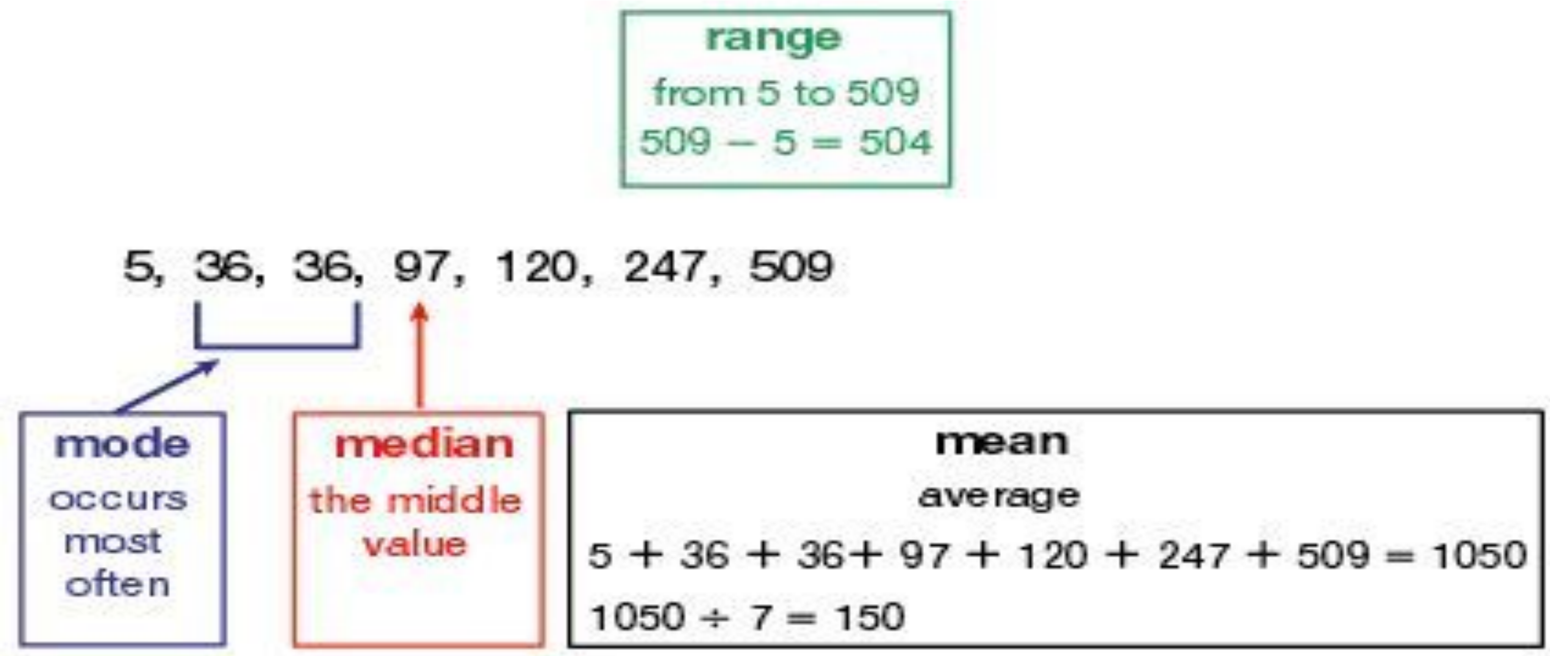
Y= Sales

Based on the graph and positive correlation, you'd be tempted to say $ spent on advertising caused an increase in sales. **But hang on -** it's also possible that an increase in sales (and thus, profit) would lead to an increase in $ spent on advertising! ***Correlation between x and y does not mean x causes y; it could mean that y causes x!***

Exploratory Analysis

Summary statistics

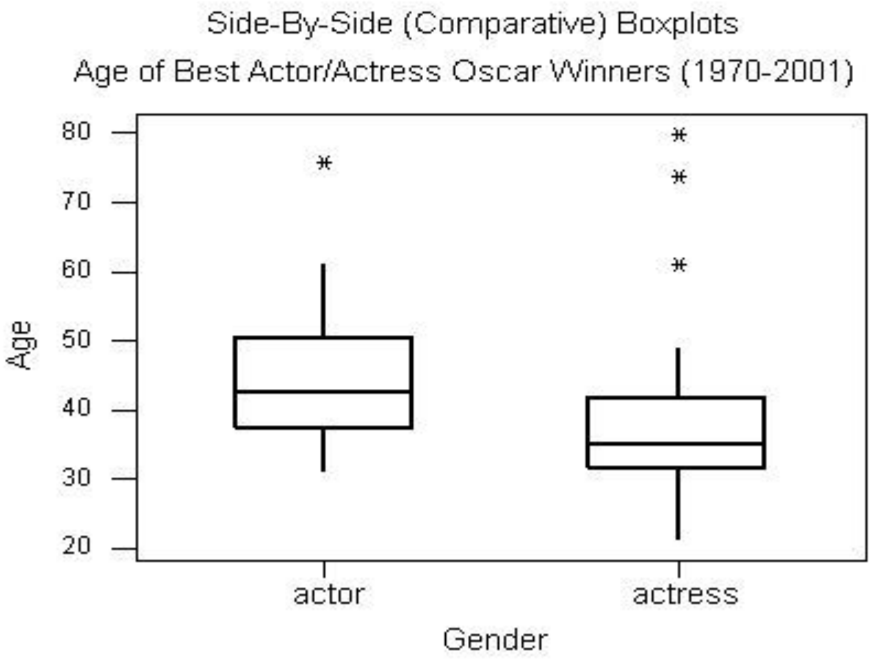
Mean, median, frequency are useful summary statistics that let you know what is in your data.



Exploratory Analysis

Boxplots

Boxplots are a useful visual depiction of certain summary statistics.



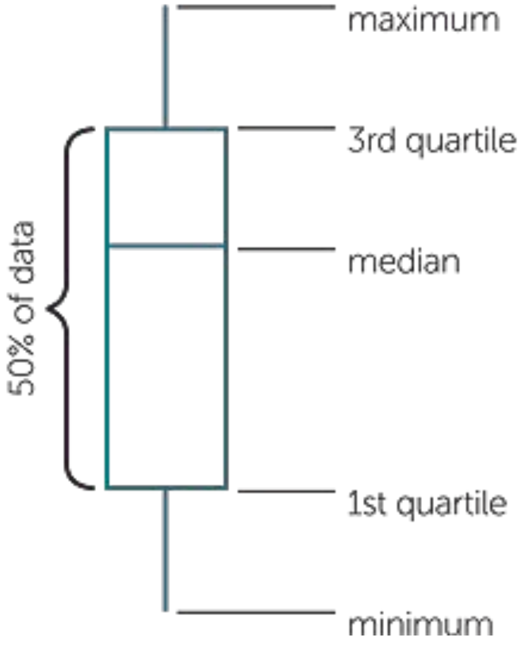


Image source: University of Florida, [Quantitative introduction to the boxplot]

Python Libraries for Data Analytics

Data Analytics

Python Libraries

Motivation for using Python for Data Analytics is the wide range of libraries

Source: https://www.pngwing.com/en/free-png-tsptz

Data Analytics

Python Libraries

- **NumPy**: NumPy supports n-dimensional arrays and provides numerical computing tools. It is useful for Linear algebra.
- **Pandas**: Pandas provides functions to handle missing data, perform mathematical operations, and manipulate the data.
- **Matplotlib**: Matplotlib library is commonly used for plotting data points and creating interactive visualizations of the data.

Data Analytics

Python Libraries

- **Seaborn**: a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.
- **SciPy**: SciPy library is used for scientific computing. It contains modules for optimization, linear algebra, integration, interpolation, special functions, signal and image processing.
- **Scikit-Learn**: Scikit-Learn library has features that allow you to build regression, classification, and clustering models.

Data Analytics to find insights on Financial Inclusion in Africa

a. Define the research question
b. Data Validation and Cleaning
c. Exploratory Analysis

Step 1: Define Research Question

Financial inclusion remains one of the main obstacles to economic and human development in Africa. For example, across Kenya, Rwanda, Tanzania, and Uganda only 9.1 million adults (or 14% of adults) have access to or use a commercial bank account.

How can we predict using a machine learning model which individuals are most likely to have or use a bank account?

Source: Financial Inclusion in Africa

Step 2: Data Validation and Cleaning

- Features of the dataset
  - Loading the dataset
  - Inspect the dataset size and data type
  - Checking for missing variables
  - Clean the data
  - Target variable (person having bank account, Yes=1, No=0)

Step 3: Exploratory Data Analysis

Finding insights before modeling

- Distribution of Target Variable
- Summary Statistics of the dataset
- Explore the gender distribution of the individuals in the countries
- Explore distribution of respondents by country
- Explore the locations of the individuals
- Explore the access to cellphones

Feature Engineering

- Encode categorical features

Coding Example Data Analytics in Python

- Financial Inclusion in Africa

Goal(s) for Coding walkthrough

- Hands-on Python Libraries for Data Analytics
  - NumPy
  - Pandas
  - Matplotlib and
  - Seaborn
- Exploratory Data Analysis (EDA)
  - Getting insights on Financial Inclusion Data
  - Visualizing the results
- Link to Jupyter notebooks

Further exploration

- Slides and Colab notebook are made available on GitHub
- Pandas documentation
- SciPy Lecture Notes

Thank you.

SOL PLAATJE
UNIVERSITY

Contact

@dmachuve

@nakadori