TRAINING PROGRAM IN ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

Instructor:

Rene Kabagamba
Arnaud Kayonga

Reference

Python Data Science Handbook

https://github.com/jakevdp/PythonDataScienceHandbook/tree/master

Chapter 02.06 - 02.09: Numpy continued

Chapter 03.01 - 03.13: Pandas

Classroom Work

1. Form groups & review chapter 03.03 - 03.08 of the Python Data Science Handbook.

Homework 8

Download the following data: https://datacatalog.worldbank.org/search/dataset/0064005/Rwanda-Maize-Production-2018---Table-Extract

On Rwanda Maize Production by district in 2018 and follow instructions for loading and analyzing the data.

Homework 8

Goal: This assignment uses pandas to explore data on Rwandan maize production.

Data Source: You'll be using a dataset on Rwandan maize production provided by the National Institute of Statistics of Rwanda (NISR).

Steps:

Import libraries and load data:

1. Start by importing pandas and any other libraries you might need (e.g., matplotlib for visualizations).
2. Assume the data is provided in a CSV file named "maize_production_2018.csv". Use pd.read_csv() to load the data into a pandas dataframe.

Note: You will have to edit the csv to remain with only the desired table with the data.

Homework 8

Explore the data:

1. Check basic information about the dataframe using df.info(), df.head(), and df.tail().
2. Identify the columns containing information you want to analyze (e.g., district names, maize production statistics).

Homework 8

Data Cleaning (Optional):

There might be missing values (indicated by NaN) in the data. You can explore them using df.isnull().sum(). Handle missing values if needed (e.g., remove rows, impute values).

Note: Even though on this one you might not have rows with missing data, though it’s a good practice to include this check.

Homework 8

Exploratory Analysis:

1. Calculate descriptive statistics for maize production across districts (e.g., mean, median, standard deviation) using df.describe().
2. Identify the top maize producing districts. You can achieve this by sorting the dataframe by production value.
3. Analyze distribution of maize production. Create histograms or boxplots to visualize the spread of production values.
4. Visualize maize production on a map of Rwanda (using libraries like geopandas).

Homework 8

Write-up:

1. Summarize your findings. Which districts produce the most maize?
2. Are there any outliers in terms of production?

You can do this within your notebooks at the end, a section called Findings.

Submit your notebook the usual way by pushing it to your github repo.

Thank You!