# Workshop - Jupyter Notebook Mastery

We have a dataset which contains data presenting a view of education on a global scale. Our task is to explore the data and set out possible problems, if any.

## Explore the metadata

To understand the data we have, we should try and find as many metadata as possible. Explore the metadata and find some datapoints which are interesting for you.



## Read our data

Read our data into a Pandas DataFrame.

Note: You might need to use **encoding=’latin-1’**.

## Fix columns

Have a look at the columns in the dataset.

Replace whitespace with underscore (“\_”)

## Drop unnecessary data

Drop the Latitude and Longitude columns since we are not going to use them.

## Derive basic statistics

Observe the **data type** of all the columns in the dataset.

If most of them are **ints** and **floats**, derive basic statistics about the dataset using the **.describe()** method.

## Find number of rows and columns

Present the number of rows and number of columns in the dataset in the following format: "Number of rows: {rows}, number of columns: {columns}"

## Find missing values

Check for missing values. Are there any?

On separate lines, print the null rate of every column that contains null values in the following format:

"{column1} null rate: {null\_rate1}"

"{column2} null rate: {null\_rate2}"

## Deal with missing values

**Replace** missing values with the most frequently met value in each respective column (**mode**).

## Find duplicates

Check for duplicates. Are there any?

## Create a new DataFrame

Create a new DataFrame which contains **all the completion rates (average)** by country.

Create a **new column** in the Dataframe, which contains the average completion rates per country.

## Top 10 and Bottom 10 countries

Plot a bar plot of the top 10 countries and bottom 10 countries by average completion rate.

Note: You can try using the **.plot(kind=’bar’)** method from Pandas.

## Columns with completion rates

Create a variable which stores all columns names of columns which contain information about the **completion rates** of **male** observations.

Do the same with **females**.

## Mean of male completion rates

Find what is the mean completion rate for every male completion rate column.

## Discover outliers

Investigate why the mean completion rates are so low.

Hint: Use **.value\_counts()**

Create a **scatter plot** of the results.

## Numerical columns

Create a new DataFrame which consists only of **numerical** columns from the original dataset.

## Correlation between variables

Find what is the **correlation coefficient** between all the columns in your new numerical dataframe.

Plot a **heatmap** of the correlation coefficients.

## Unemployment drivers

What are the most correlated variables with the unemployment measurement?

Note: Consider using **absolute** values.

## Group unemployment by country

Group your dataset by **country**, finding the **mean** of the **unemployment** **rate**.

## Top 15 countries by unemployment

Find the top 15 countries by unemployment.

Plot a **horizontal bar plot**.

**Figure size** should be 9 by 7.

**Figure dpi** should be 300.

**First three bars** should be with the color “#**32a852**”

**The rest of the bars** should be with the color “#**9ad6aa**”.

## Birth rate and enrollment

Plot a **scatter plot** of the birth rate and the two enrollment variables.

Place a **legend** in the **upper left corner**.