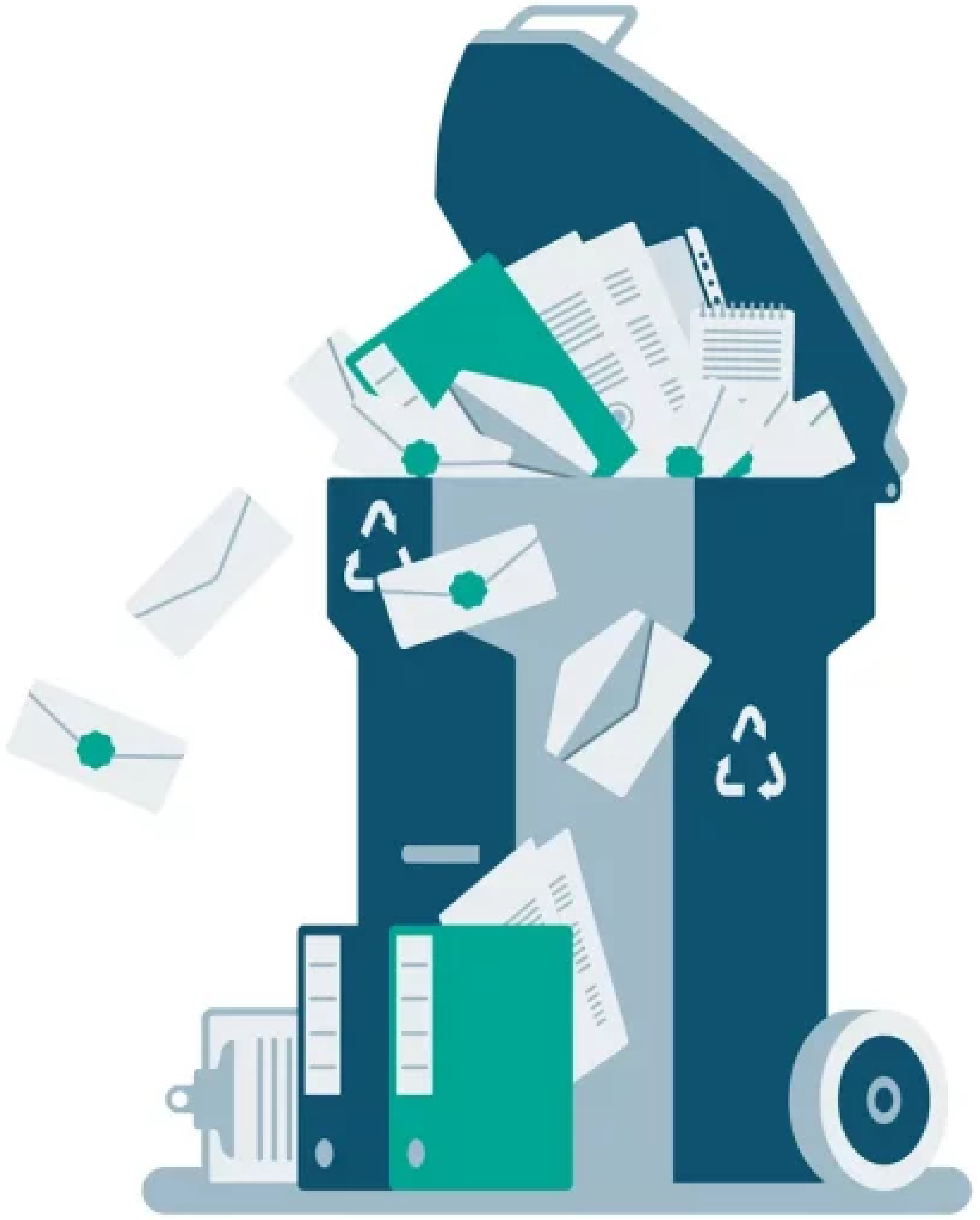
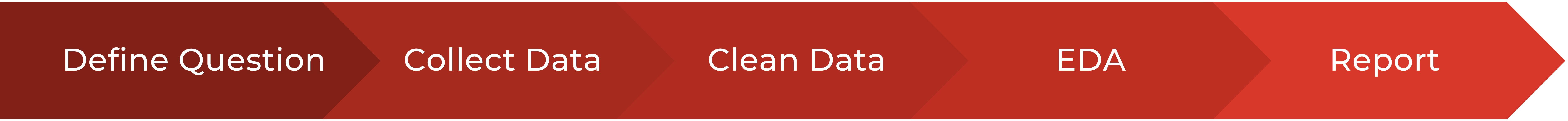
Hello pandas

GARBAGE IN, GARBAGE OUT





Define Question

Collect Data

Clean Data

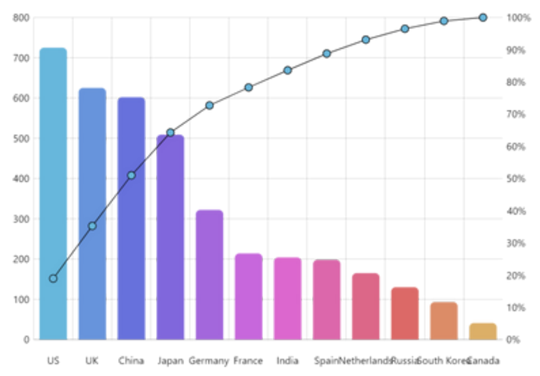
EDA

Report

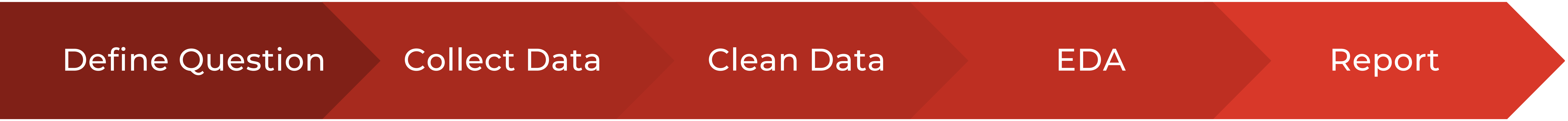
How to increase our customer engagement?

Web scraping Database

Duplicated values,
Structural errors,
Outliers,
Missing data,
Sort and organize data,
irrelevant data,
Validate data,
Transform data



Exploratory Data Analysis



Define Question

Collect Data

Clean Data

EDA

Report

Python,
Oracle,
MySQL,
SQLServer,...

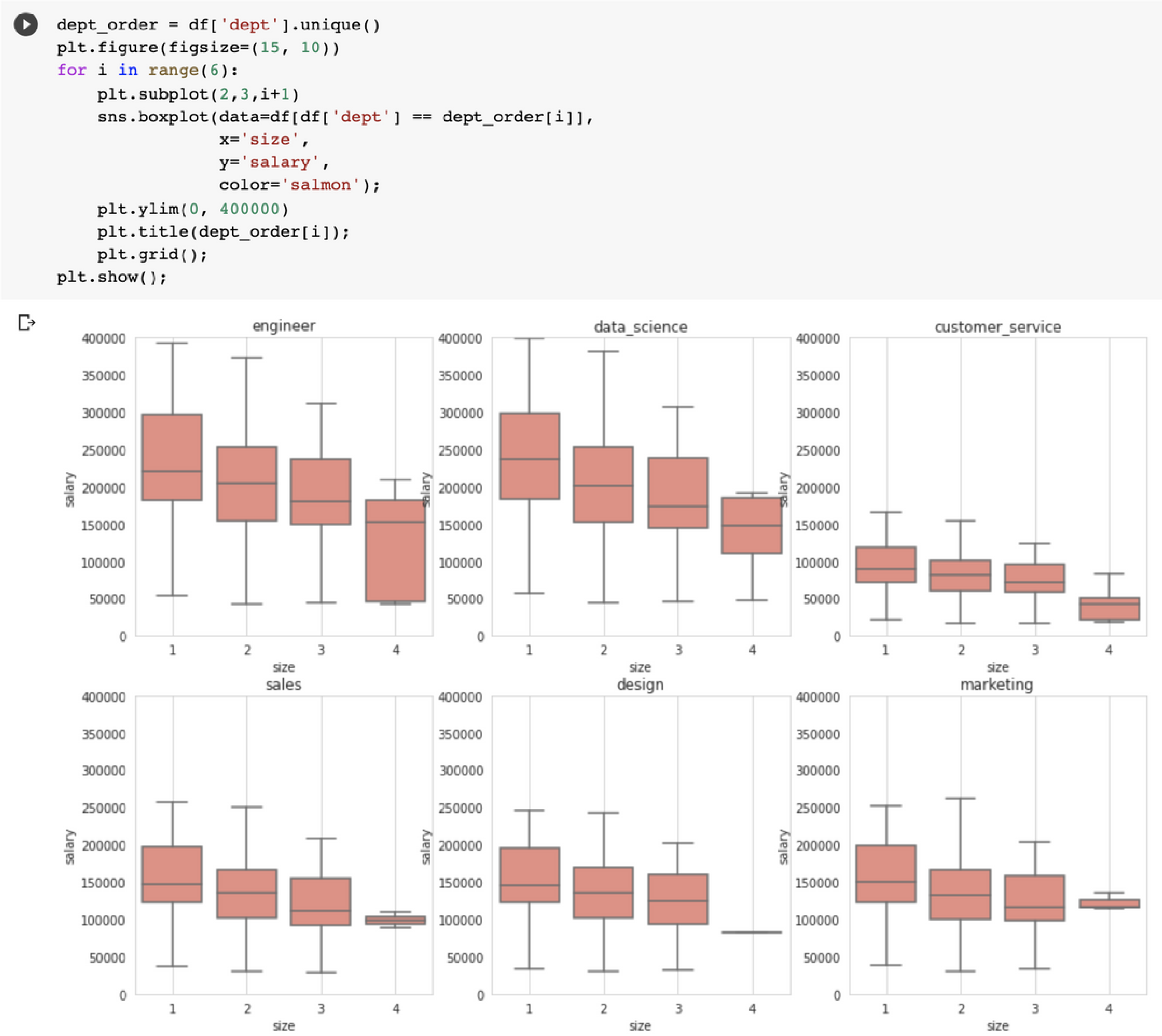
Pandas

Pandas

Matplotlib
Seaborn

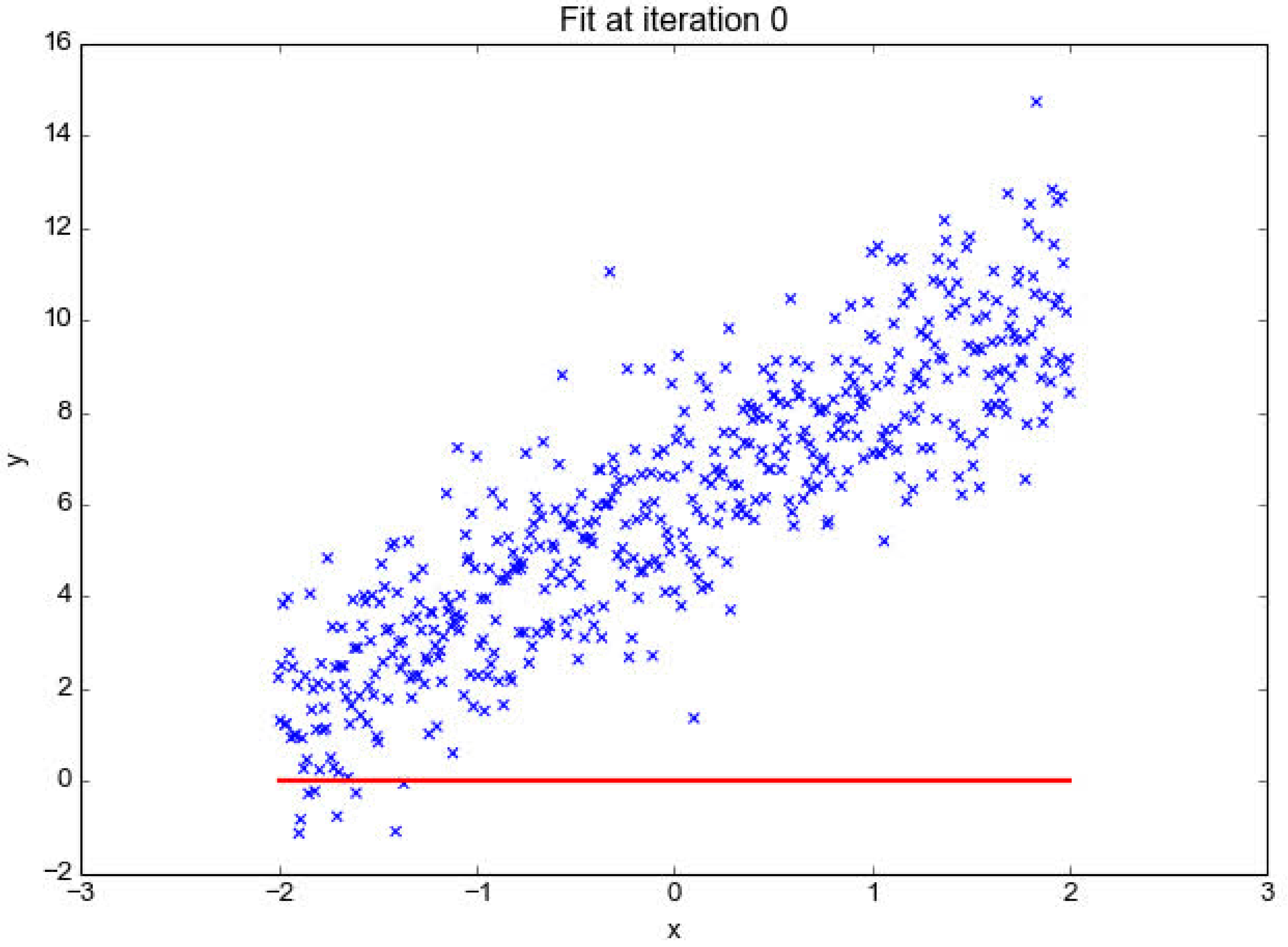
PowerBI,
Tableau,
Google Data Studio

Why Python / Pandas?



```
dept_order = df['dept'].unique()
plt.figure(figsize=(15, 10))
for i in range(6):
    plt.subplot(2,3,i+1)
    sns.boxplot(data=df[df['dept'] == dept_order[i]],
                x='size',
                y='salary',
                color='salmon');
    plt.ylim(0, 400000)
    plt.title(dept_order[i]);
    plt.grid();
plt.show();
```

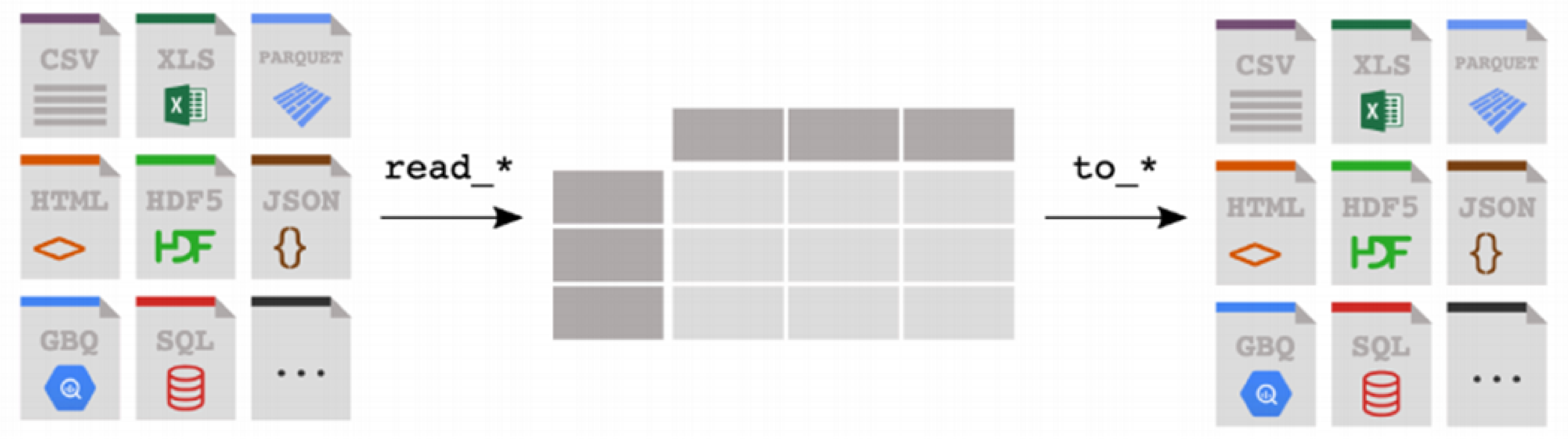
immediate output



modelling / machine learning

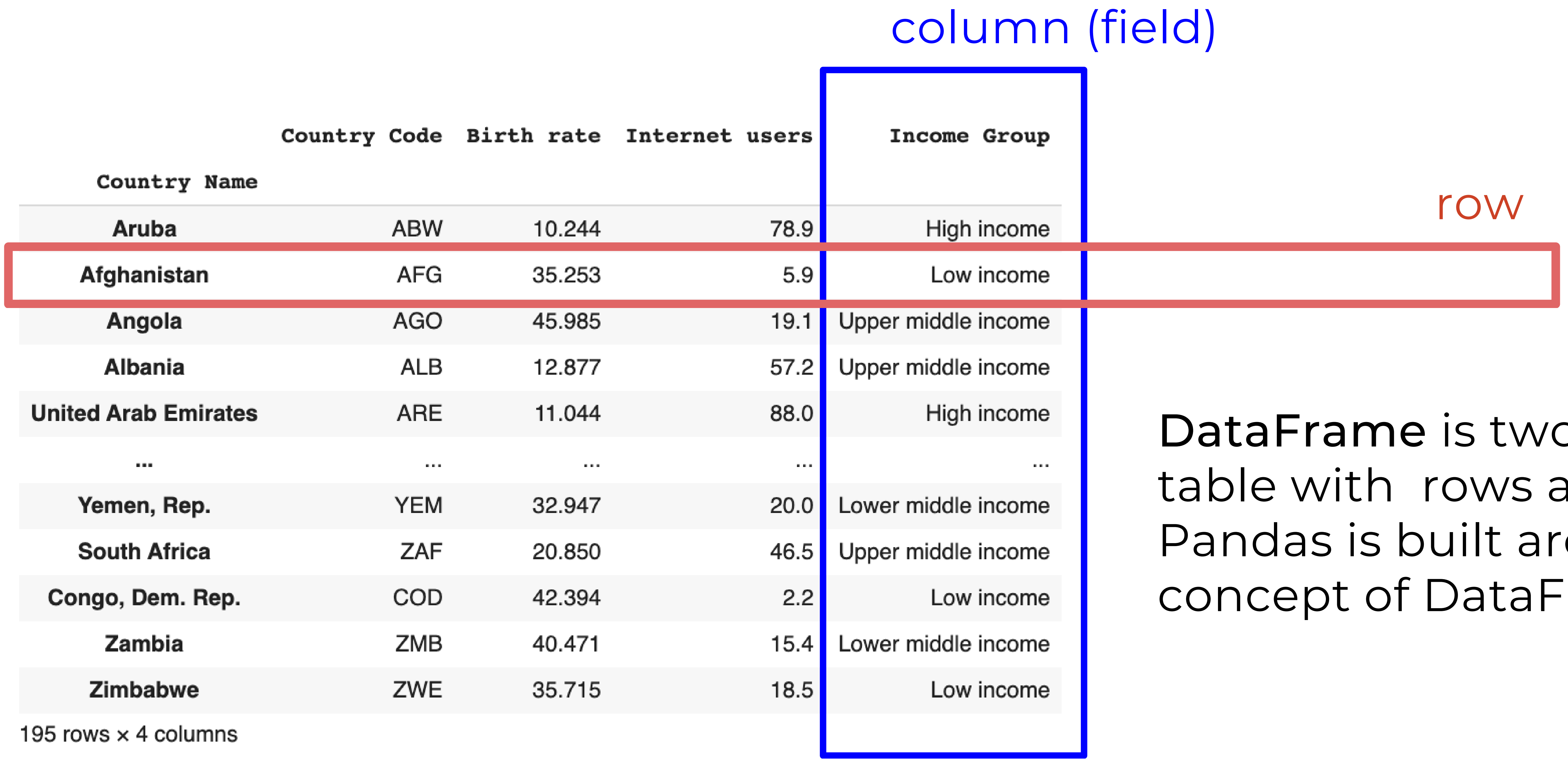
WHY pandas ?

- **Pandas** is an open-source library used for data analysis.
- It's fast, powerful, flexible, and easy to use
- It supports many different types of data formats



NOT THIS PANDA :))

Pandas Component - DataFrame

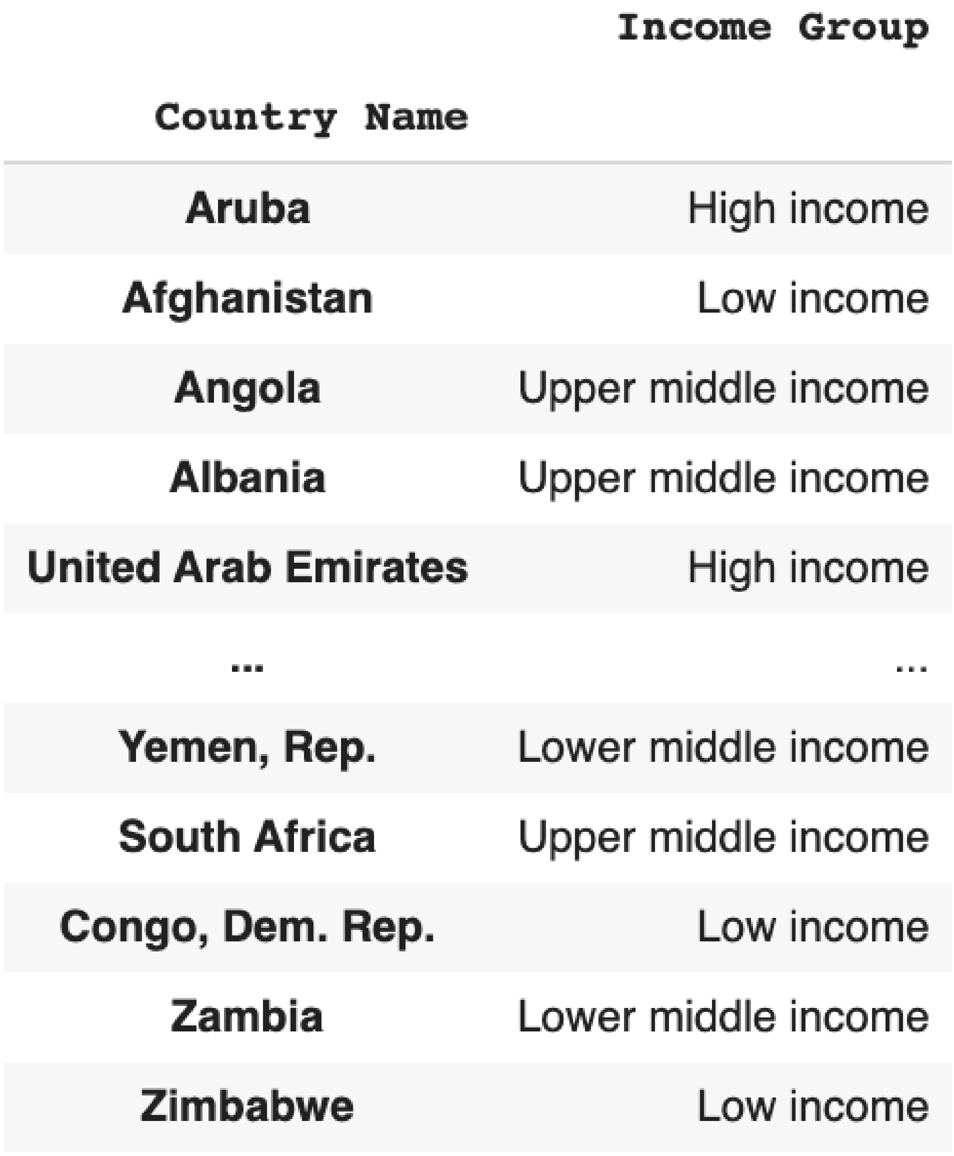


	Country Code	Birth rate	Internet users	Income Group
Country Name				
Aruba	ABW	10.244	78.9	High income
Afghanistan	AFG	35.253	5.9	Low income
Angola	AGO	45.985	19.1	Upper middle income
Albania	ALB	12.877	57.2	Upper middle income
United Arab Emirates	ARE	11.044	88.0	High income
...	...	...	...	...
Yemen, Rep.	YEM	32.947	20.0	Lower middle income
South Africa	ZAF	20.850	46.5	Upper middle income
Congo, Dem. Rep.	COD	42.394	2.2	Low income
Zambia	ZMB	40.471	15.4	Lower middle income
Zimbabwe	ZWE	35.715	18.5	Low income

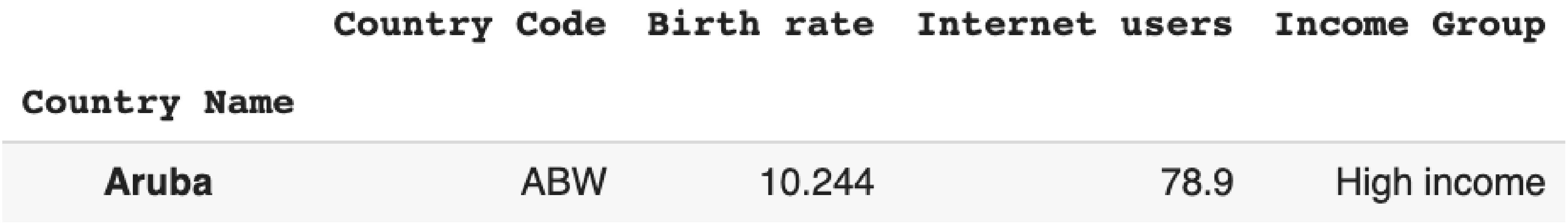
195 rows × 4 columns

DataFrame is two dimensional table with rows and columns. Pandas is built around the concept of DataFrame.

Series is one-dimensional array with axis label. One row (one column) is a Series. **One Series has only one data type.**



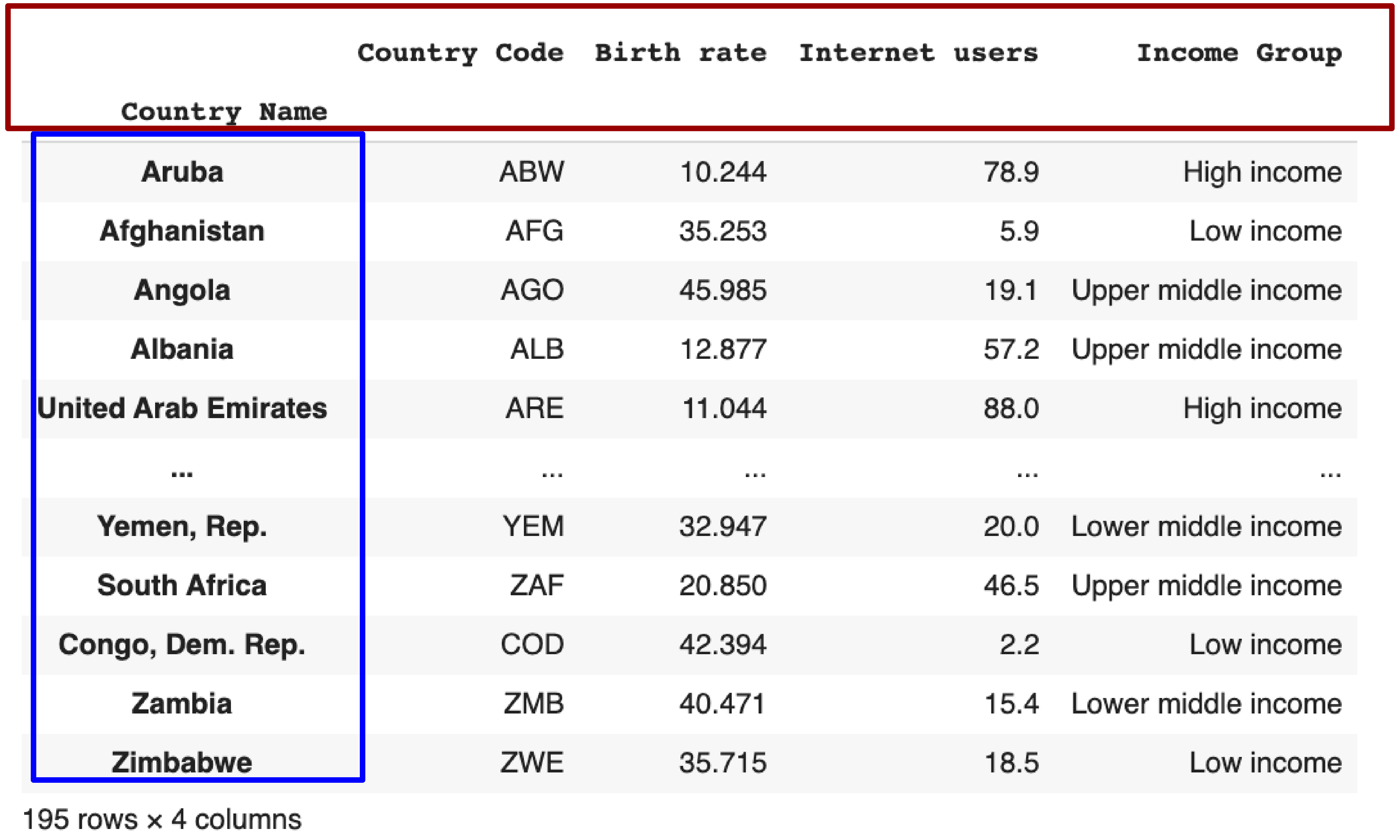
	Income Group
Country Name	
Aruba	High income
Afghanistan	Low income
Angola	Upper middle income
Albania	Upper middle income
United Arab Emirates	High income
...	...
Yemen, Rep.	Lower middle income
South Africa	Upper middle income
Congo, Dem. Rep.	Low income
Zambia	Lower middle income
Zimbabwe	Low income



	Country Code	Birth rate	Internet users	Income Group
Country Name				
Aruba	ABW	10.244	78.9	High income

Pandas Component - Series

Index is like an address, that’s how any data point across the DataFrame or Series can be accessed. Rows and columns both have indexes.



	Country Code	Birth rate	Internet users	Income Group
Country Name				
Aruba	ABW	10.244	78.9	High income
Afghanistan	AFG	35.253	5.9	Low income
Angola	AGO	45.985	19.1	Upper middle income
Albania	ALB	12.877	57.2	Upper middle income
United Arab Emirates	ARE	11.044	88.0	High income
...	...	...	...	...
Yemen, Rep.	YEM	32.947	20.0	Lower middle income
South Africa	ZAF	20.850	46.5	Upper middle income
Congo, Dem. Rep.	COD	42.394	2.2	Low income
Zambia	ZMB	40.471	15.4	Lower middle income
Zimbabwe	ZWE	35.715	18.5	Low income

195 rows × 4 columns

Pandas Component - Index

Let's **practice!**