df.tail() ■ df.shape ✓ • Basic ops on columns 🗸 lacktriangleright Different ways of accessing cols \cV ■ Check for Unique values ✓ ■ Rename column ✓ ■ Deleting col ✓ ■ Creating new cols ✓ • Basic ops on rows Implicit/explicit index df.index Indexing in series Slicing in series ■ loc/iloc Indexing/Slicing in dataframe Adding a row Deleting a row Check for duplicates Working with both rows and columns **Installation of pandas** Importing pandas In [1]: !pip install pandas Requirement already satisfied: pandas in /Users/nikhilsanghi/opt/anaconda3/lib/python3.9/site-packages (1.3.4) Requirement already satisfied: python-dateutil>=2.7.3 in /Users/nikhilsanghi/opt/anaconda3/lib/python3.9/site-packages (from pandas) (2.8.2) Requirement already satisfied: pytz>=2017.3 in /Users/nikhilsanghi/opt/anaconda3/lib/python3.9/site-packages (from pandas) (2021.3) Requirement already satisfied: numpy>=1.17.3 in /Users/nikhilsanghi/opt/anaconda3/lib/python3.9/site-packages (from pandas) (1.20.3) Requirement already satisfied: six>=1.5 in /Users/nikhilsanghi/opt/anaconda3/lib/python3.9/site-packages (from python-dateutil>=2.7.3->pandas) (1.16.0) In [2]: import numpy as np import pandas as pd Importing the dataset In [47]: df=pd.read_csv("/Users/nikhilsanghi/Downloads/01_dsml-course-main-live/batches/2_Sept_Beg_Tue_Oct_Beg_Tue/07_Pandas_1/mckinsey.csv") country year population continent life_exp Out[47]: gdp_cap **0** Afghanistan 1952 8425333 28.801 779.445314 Asia 1 Afghanistan 1957 9240934 30.332 820.853030 **2** Afghanistan 1962 10267083 31.997 853.100710 **3** Afghanistan 1967 11537966 34.020 836.197138 4 Afghanistan 1972 13079460 Asia 36.088 739.981106 706.157306 Zimbabwe 1987 9216418 Africa 62.351 Zimbabwe 1992 10704340 60.377 693.420786 1700 Africa Zimbabwe 1997 11404948 46.809 792.449960 Africa 11926563 39.989 672.038623 Zimbabwe 2002 Africa 12311143 Zimbabwe 2007 43.487 469.709298 1704 rows × 6 columns In [4]: type(df) pandas.core.frame.DataFrame Out[4]: Dataframe/Series In [5]: # 1. row oriented approach df1=pd.DataFrame([["Afghanistan", 1952, 8425333, "Asia", 28.801, 779.445314], ["Afghanistan", 1957, 9240934, "Asia", 30.332, 820.853030]], columns=["country", "year", "population", "continent", "life_exp", "gdp_cap"]) df1 country year population continent life_exp Out[7]: gdp_cap **0** Afghanistan 1952 8425333 28.801 779.445314 **1** Afghanistan 1957 9240934 Asia 30.332 820.853030 In []: # 2. column oriented approach In [9]: df2=pd.DataFrame({ "country":["Afghanistan", "Afghanistan"], "year":[1952,1957], "population":[8425333,9240934], "continent":["Asia", "Asia"], "life_exp":[28.801,30.332], "gdp_cap": [779.445314,820.853030], df2 population continent life_exp Out[9]: gdp_cap country year **0** Afghanistan 1952 8425333 28.801 779.445314 1 Afghanistan 1957 9240934 30.332 820.853030 Asia **Basic ops on a DataFrame** df.info() In [15]: df.info() <class 'pandas.core.frame.DataFrame'> RangeIndex: 1704 entries, 0 to 1703 Data columns (total 6 columns): Column Non-Null Count Dtype 0 country 1704 non-null object 1704 non-null int64 1 year population 1704 non-null int64 continent 1704 non-null object life_exp 1704 non-null float64 1704 non-null float64 gdp_cap dtypes: float64(2), int64(2), object(2) memory usage: 80.0+ KB df.describe() In [16]: df.describe() Out[16]: population life_exp gdp_cap count 1704.00000 1.704000e+03 1704.000000 1704.000000 mean 1979.50000 2.960121e+07 59.474439 7215.327081 17.26533 1.061579e+08 12.917107 9857.454543 min 1952.00000 6.001100e+04 23.599000 241.165876 **25**% 1965.75000 2.793664e+06 48.198000 1202.060309 60.712500 **50%** 1979.50000 7.023596e+06 3531.846988 **75**% 1993.25000 1.958522e+07 70.845500 9325.462346 82.603000 113523.132900 max 2007.00000 1.318683e+09 In [17]: df.describe(include="object") Out[17]: country continent count 1704 1704 5 unique 142 top Afghanistan Africa 12 624 freq In [18]: df.describe(include="all") Out[18]: country population continent life_exp gdp_cap 1704 1704.000000 1704 1704.00000 1.704000e+03 1704.000000 count NaN NaN NaN NaN unique top Afghanistan NaN NaN Africa NaN NaN NaN NaN 624 NaN NaN freq NaN 1979.50000 2.960121e+07 59.474439 NaN 7215.327081 mean std 17.26533 1.061579e+08 NaN 12.917107 9857.454543 NaN 1952.00000 6.001100e+04 23.599000 241.165876 min NaN **25**% NaN 1965.75000 2.793664e+06 NaN 48.198000 1202.060309 **50**% NaN 1979.50000 7.023596e+06 60.712500 3531.846988 NaN **75**% NaN 1993.25000 1.958522e+07 70.845500 9325.462346 NaN NaN 2007.00000 1.318683e+09 82.603000 113523.132900 max NaN df.head() In [10] df.head() country year population continent life_exp Out[10]: gdp_cap 28.801 779.445314 **0** Afghanistan 1952 8425333 Asia 1 Afghanistan 1957 9240934 Asia 30.332 820.853030 2 Afghanistan 1962 10267083 Asia 31.997 853.100710 **3** Afghanistan 1967 11537966 34.020 836.197138 Asia 4 Afghanistan 1972 13079460 36.088 739.981106 Asia In [11]: df.head(10) Out[11]: country year population continent life_exp gdp_cap **0** Afghanistan 1952 28.801 779.445314 8425333 Asia **1** Afghanistan 1957 9240934 30.332 820.853030 2 Afghanistan 1962 10267083 31.997 853.100710 Asia **3** Afghanistan 1967 11537966 Asia 34.020 836.197138 4 Afghanistan 1972 13079460 36.088 739.981106 Asia **5** Afghanistan 1977 14880372 38.438 786.113360 Asia 39.854 978.011439 6 Afghanistan 1982 12881816 Asia **7** Afghanistan 1987 13867957 40.822 852.395945 Asia 8 Afghanistan 1992 16317921 41.674 649.341395 Asia **9** Afghanistan 1997 22227415 41.763 635.341351 Asia df.tail() In [12]: df.tail() Out[12]: country year population continent life_exp gdp_cap **1699** Zimbabwe 1987 9216418 62.351 706.157306 Africa **1700** Zimbabwe 1992 60.377 693.420786 10704340 Africa **1701** Zimbabwe 1997 11404948 Africa 46.809 792.449960 **1702** Zimbabwe 2002 11926563 Africa 39.989 672.038623 **1703** Zimbabwe 2007 12311143 Africa 43.487 469.709298 In [13]: df.tail(10) gdp_cap Out[13]: country year population continent life_exp **1694** Zimbabwe 1962 4277736 Africa 52.358 527.272182 **1695** Zimbabwe 1967 4995432 Africa 53.995 569.795071 **1696** Zimbabwe 1972 5861135 55.635 799.362176 Africa **1697** Zimbabwe 1977 6642107 57.674 685.587682 Africa **1698** Zimbabwe 1982 7636524 60.363 788.855041 Africa **1699** Zimbabwe 1987 9216418 Africa 62.351 706.157306 **1700** Zimbabwe 1992 10704340 60.377 693.420786 Africa **1701** Zimbabwe 1997 11404948 Africa 46.809 792.449960 39.989 672.038623 **1702** Zimbabwe 2002 11926563 **1703** Zimbabwe 2007 12311143 43.487 469.709298 df.shape In [14]: df.shape (1704, 6)Out[14]: **Basic ops on columns Accessing Column Names** In [30]: df.keys() Index(['country', 'year', 'population', 'continent', 'life_exp', 'gdp_cap'], dtype='object') Out[30]: In [31]: df.columns Index(['country', 'year', 'population', 'continent', 'life_exp', 'gdp_cap'], dtype='object') Different ways of accessing cols In [19]: country year population continent life_exp Out[19]: gdp_cap **0** Afghanistan 1952 28.801 779.445314 8425333 1 Afghanistan 1957 9240934 30.332 820.853030 Asia 10267083 31.997 853.100710 **2** Afghanistan 1962 **3** Afghanistan 1967 11537966 Asia 34.020 836.197138 36.088 739.981106 4 Afghanistan 1972 13079460 Zimbabwe 1987 9216418 62.351 706.157306 1699 Africa 10704340 60.377 693.420786 1700 Zimbabwe 1992 Africa 11404948 46.809 792.449960 Zimbabwe 1997 11926563 1702 Zimbabwe 2002 39.989 672.038623 Africa 43.487 469.709298 Zimbabwe 2007 12311143 Africa 1704 rows × 6 columns df["country"] # method 1 # recommended Afghanistan Out[20]: Afghanistan Afghanistan 3 Afghanistan Afghanistan 1699 Zimbabwe 1700 Zimbabwe Zimbabwe 1701 1702 Zimbabwe 1703 Zimbabwe Name: country, Length: 1704, dtype: object type(df["country"]) pandas.core.series.Series In [23]: df[["country", "year"]] Out[23]: country year **0** Afghanistan 1952 **1** Afghanistan 1957 2 Afghanistan 1962 **3** Afghanistan 1967 4 Afghanistan 1972 Zimbabwe 1987 Zimbabwe 1992 Zimbabwe 1997 Zimbabwe 2002 Zimbabwe 2007 1704 rows × 2 columns # method 2 # not recommended df.country Afghanistan Out[24]: Afghanistan Afghanistan Afghanistan Afghanistan 1699 Zimbabwe 1700 Zimbabwe 1701 Zimbabwe 1702 Zimbabwe 1703 Zimbabwe Name: country, Length: 1704, dtype: object # df.max # df.country name # df.1 # df.country, year In []: In []: Check for Unique values In [26]: df["continent"].unique() array(['Asia', 'Europe', 'Africa', 'Americas', 'Oceania'], dtype=object) Out[26]: df["continent"].nunique() df["continent"].value_counts() Africa 624 Out[28]: 396 Europe 360 Americas 300 24 Oceania Name: continent, dtype: int64 Rename column In [29]: df.head() gdp_cap Out[29]: country year population continent life_exp **0** Afghanistan 1952 8425333 28.801 779.445314 1 Afghanistan 1957 9240934 30.332 820.853030 Asia 2 Afghanistan 1962 10267083 31.997 853.100710 Asia **3** Afghanistan 1967 11537966 Asia 34.020 836.197138 4 Afghanistan 1972 13079460 36.088 739.981106 Asia In [36]: #method 1 df.rename({"country":"Country"}, axis=1) Out[36]: Country year population continent life_exp gdp_cap **0** Afghanistan 1952 8425333 28.801 779.445314 Asia 1 Afghanistan 1957 9240934 30.332 820.853030 Asia 2 Afghanistan 1962 10267083 Asia 31.997 853.100710 3 Afghanistan 1967 11537966 Asia 34.020 836.197138 4 Afghanistan 1972 13079460 36.088 739.981106 Asia 1699 Zimbabwe 1987 9216418 Africa 62.351 706.157306 10704340 Zimbabwe 1992 Africa 60.377 693.420786 Zimbabwe 1997 11404948 Africa 46.809 792.449960 1702 Zimbabwe 2002 11926563 Africa 39.989 672.038623 1703 Zimbabwe 2007 12311143 Africa 43.487 469.709298 1704 rows × 6 columns In [37]: df.rename(columns={"country":"Country"}) Out[37]: Country year population continent life_exp gdp_cap **0** Afghanistan 1952 8425333 28.801 779.445314 Asia 1 Afghanistan 1957 9240934 30.332 820.853030 Asia **2** Afghanistan 1962 10267083 Asia 31.997 853.100710 3 Afghanistan 1967 11537966 34.020 836.197138 Asia 4 Afghanistan 1972 13079460 36.088 739.981106 Asia 62.351 706.157306 Zimbabwe 1987 9216418 Africa Zimbabwe 1992 10704340 Africa 60.377 693.420786 11926563 1702 Zimbabwe 2002 Africa 39.989 672.038623 12311143 43.487 469.709298 Zimbabwe 2007 1704 rows × 6 columns In [38] df.head() country year population continent life_exp gdp_cap Out[38]: **0** Afghanistan 1952 8425333 28.801 779.445314 Asia **1** Afghanistan 1957 9240934 Asia 30.332 820.853030 **2** Afghanistan 1962 10267083 31.997 853.100710 Asia **3** Afghanistan 1967 11537966 Asia 34.020 836.197138 36.088 739.981106 4 Afghanistan 1972 13079460 Asia In [39]: # Method 1 (preferrable) df.rename({"country":"Country"}, axis=1, inplace=True) In [40]: df.head() gdp_cap Out[40]: Country year population continent life_exp 8425333 28.801 779.445314 **0** Afghanistan 1952 Asia 30.332 820.853030 1 Afghanistan 1957 9240934 Asia **2** Afghanistan 1962 10267083 Asia 31.997 853.100710 11537966 34.020 836.197138 3 Afghanistan 1967 Asia 4 Afghanistan 1972 13079460 36.088 739.981106 Asia In [41]: # Method 2 df=df.rename({"year":"Year"},axis=1) In [42] df.head() Country Year population continent life_exp Out[42]: gdp_cap **0** Afghanistan 1952 8425333 28.801 779.445314 9240934 30.332 820.853030 **1** Afghanistan 1957 Asia 10267083 31.997 853.100710 **2** Afghanistan 1962 Asia **3** Afghanistan 1967 11537966 Asia 34.020 836.197138 4 Afghanistan 1972 36.088 739.981106 13079460 In [43]: # Method 1 (preferrable) df.rename(columns={"population":"Population"},inplace=True) In [44]: df.head() Out[44]: Country Year Population continent life_exp gdp_cap **0** Afghanistan 1952 8425333 28.801 779.445314 Asia 1 Afghanistan 1957 9240934 Asia 30.332 820.853030 **2** Afghanistan 1962 10267083 31.997 853.100710 Asia 11537966 34.020 836.197138 **3** Afghanistan 1967 Asia **4** Afghanistan 1972 13079460 Asia 36.088 739.981106 In [45]: # Method 2 df=df.rename(columns={"life_exp":"Life_exp"}) In [46]: df.head() Country Year Population continent Life_exp Out[46]: gdp_cap 28.801 779.445314 **0** Afghanistan 1952 8425333 Asia 1 Afghanistan 1957 9240934 30.332 820.853030 Asia **2** Afghanistan 1962 10267083 31.997 853.100710 Asia 3 Afghanistan 1967 11537966 Asia 34.020 836.197138 36.088 739.981106 4 Afghanistan 1972 13079460 Asia In [52]: df.rename(columns={"country":"Country", "year":"Year", "continent": "Continent", "population": "Population", "life_exp":"Life_exp", "gdp_cap":"Gdp_cap" },inplace=True) In [53]: df.head() Country Year Population Continent Life_exp Gdp_cap Out[53]: 28.801 779.445314 **0** Afghanistan 1952 8425333 Asia **1** Afghanistan 1957 9240934 30.332 820.853030 Asia **2** Afghanistan 1962 10267083 31.997 853.100710 Asia 3 Afghanistan 1967 11537966 34.020 836.197138 Asia 36.088 739.981106 4 Afghanistan 1972 13079460 Asia Deleting col In [54]: df.head() Out[54]: Country Year Population Continent Life_exp Gdp_cap **0** Afghanistan 1952 8425333 Asia 28.801 779.445314 1 Afghanistan 1957 9240934 Asia 30.332 820.853030 **2** Afghanistan 1962 10267083 Asia 31.997 853.100710 **3** Afghanistan 1967 11537966 34.020 836.197138 Asia 4 Afghanistan 1972 36.088 739.981106 13079460 Asia In [61]: # Method 1 df.drop(["Continent"],axis=1) Out[61]: Country Year Population Life_exp Gdp_cap 28.801 779.445314 **0** Afghanistan 1952 8425333 30.332 820.853030 1 Afghanistan 1957 9240934 **2** Afghanistan 1962 10267083 31.997 853.100710 **3** Afghanistan 1967 11537966 34.020 836.197138 **4** Afghanistan 1972 13079460 36.088 739.981106 9216418 62.351 706.157306 1699 Zimbabwe 1987 1700 Zimbabwe 1992 10704340 60.377 693.420786 11404948 Zimbabwe 1997 46.809 792.449960 11926563 39.989 672.038623 1702 Zimbabwe 2002 Zimbabwe 2007 12311143 43.487 469.709298 1704 rows × 5 columns In [62]: # Method 2 df.drop(columns=["Population"]) Country Year Continent Life_exp Gdp_cap Out[62]: **0** Afghanistan 1952 Asia 28.801 779.445314 1 Afghanistan 1957 30.332 820.853030 31.997 853.100710 2 Afghanistan 1962 Asia 3 Afghanistan 1967 34.020 836.197138 Asia 4 Afghanistan 1972 Asia 36.088 739.981106 Zimbabwe 1987 Africa 62.351 706.157306 1699 60.377 693.420786 1700 Zimbabwe 1992 Africa Zimbabwe 1997 Africa 46.809 792.449960 1701 Zimbabwe 2002 Africa 39.989 672.038623 1703 Zimbabwe 2007 Africa 43.487 469.709298 1704 rows × 5 columns In [65]: df.drop(["Continent", "Year"], axis=1, inplace=True) Country Population Life_exp Gdp_cap Out[65]: **0** Afghanistan 8425333 28.801 779.445314 9240934 **1** Afghanistan 30.332 820.853030 10267083 31.997 853.100710 2 Afghanistan **3** Afghanistan 11537966 34.020 836.197138 4 Afghanistan 13079460 36.088 739.981106 1699 Zimbabwe 9216418 62.351 706.157306 Zimbabwe 10704340 60.377 693.420786 1701 Zimbabwe 11404948 46.809 792.449960 Zimbabwe 11926563 39.989 672.038623 1703 Zimbabwe 12311143 43.487 469.709298 1704 rows × 4 columns Creating new cols In [66]: df.head() Country Year Population Continent Life_exp Gdp_cap Out[66]: **0** Afghanistan 1952 8425333 28.801 779.445314 Asia 1 Afghanistan 1957 9240934 Asia 30.332 820.853030 **2** Afghanistan 1962 10267083 31.997 853.100710 Asia 3 Afghanistan 1967 11537966 34.020 836.197138 Asia 4 Afghanistan 1972 13079460 Asia 36.088 739.981106 In [67]: df["Year+7"]=df["Year"]+7 In [68]: df.head() Out[68]: Country Year Population Continent Life_exp Gdp_cap Year+7 **0** Afghanistan 1952 8425333 Asia 28.801 779.445314 1959 1 Afghanistan 1957 9240934 30.332 820.853030 1964 Asia **2** Afghanistan 1962 10267083 Asia 31.997 853.100710 1969 **3** Afghanistan 1967 11537966 Asia 34.020 836.197138 1974 4 Afghanistan 1972 13079460 Asia 36.088 739.981106 In [69]: df["Year+Gdp_cap"]=df["Year"]+df["Gdp_cap"] df.head() Gdp_cap Year+7 Year+Gdp_cap Out[69]: Country Year Population Continent Life_exp **0** Afghanistan 1952 8425333 Asia 28.801 779.445314 2731.445314 1 Afghanistan 1957 9240934 30.332 820.853030 1964 2777.853030 Asia **2** Afghanistan 1962 10267083 Asia 31.997 853.100710 2815.100710 **3** Afghanistan 1967 11537966 34.020 836.197138 1974 2803.197138 Asia 4 Afghanistan 1972 13079460 Asia 36.088 739.981106 1979 2711.981106 In [70]: df["Year+Gdp_cap"]=df["Year"] df.head() Gdp_cap Year+7 Year+Gdp_cap Country Year Population Continent Life_exp Out[70]: **0** Afghanistan 1952 8425333 28.801 779.445314 1959 1952 Asia 1 Afghanistan 1957 9240934 30.332 820.853030 1964 1957 Asia 2 Afghanistan 1962 10267083 31.997 853.100710 1969 1962 Asia **3** Afghanistan 1967 1967 11537966 Asia 34.020 836.197138 1974 4 Afghanistan 1972 13079460 Asia 36.088 739.981106 1979 1972 In [72]: # df["Country+Gdp_cap"]=df["Country"]+df["Gdp_cap"] df.head() Gdp_cap Year+7 Year+Gdp_cap Out[72]: Country Year Population Continent Life_exp **0** Afghanistan 1952 8425333 Asia 28.801 779.445314 1959 1952 1957 **1** Afghanistan 1957 9240934 Asia 30.332 820.853030 1964 2 Afghanistan 1962 10267083 31.997 853.100710 1969 1962 Asia **3** Afghanistan 1967 11537966 34.020 836.197138 1974 1967 4 Afghanistan 1972 13079460 36.088 739.981106 1979 1972 Asia df["Country+Continent"]=df["Country"]+"-"+df["Continent"] df.head() Country Year Population Continent Life_exp Gdp_cap Year+7 Year+Gdp_cap Country+Continent Out[73]: **0** Afghanistan 1952 8425333 28.801 779.445314 1959 1952 Afghanistan-Asia Asia 1 Afghanistan 1957 9240934 30.332 820.853030 1964 1957 Afghanistan-Asia Asia **2** Afghanistan 1962 10267083 31.997 853.100710 1969 1962 Afghanistan-Asia Asia **3** Afghanistan 1967 11537966 34.020 836.197138 1974 1967 Afghanistan-Asia Asia 4 Afghanistan 1972 13079460 Asia 36.088 739.981106 1979 1972 Afghanistan-Asia In []: In []: In []: In []: **Basic ops on rows** Implicit/explicit index In []: df.index In []: Indexing in series In []: Slicing in series In [] loc/iloc In []: Indexing/Slicing in dataframe In []: Adding a row Deleting a row In []:

Check for duplicates

Working with both rows and columns

In []:

In []:

In []:

In []:

Agenda

• Installation of pandas 🗸

■ Importing pandas ✓

■ Dataframe/Series ✓

Basic ops on a DataFrame

df.info()
df.head()

■ Importing the dataset ✓