Colab: https://colab.research.google.com/drive/1fkgXd0C2Nh_hPybBxn1oV9MU4TPEVSmN?usp=sharing

```
import pandas as pd
```

```
import numpy as np
```

https://drive.google.com/file/d/1E3bwvYGf1ig32RmcYiWc0IXPN-mD_bI_/view?usp=sharing

```
!gdown 1E3bwvYGf1ig32RmcYiWc0IXPN-mD_bI_
```

```
Downloading...
From: https://drive.google.com/uc?id=1E3bwvYGf1ig32RmcYiWc0IXPN-mD_bI_
To: /content/mckinsey.csv
100% 83.8k/83.8k [00:00<00:00, 39.0MB/s]
```

```
df = pd.read_csv("mckinsey.csv")
df
```

```
type(df)
```

```
pandas.core.frame.DataFrame
```

```
df["population"]
```

```
0        8425333
1        9240934
2       10267083
3       11537966
4       13079460
          ...
1699     9216418
1700    10704340
1701    11404948
1702    11926563
1703    12311143
Name: population, Length: 1704, dtype: int64
```

```
type(df["population"])
```

```
pandas.core.series.Series
```

```
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
 1   year        1704 non-null   int64
 2   population  1704 non-null   int64
 3   continent   1704 non-null   object
 4   life_exp    1704 non-null   float64
 5   gdp_cap     1704 non-null   float64
```

```
    dtypes: float64(2), int64(2), object(2)
    memory usage: 80.0+ KB
```

```
df.head(7)
```

```
df.tail()
```

```
df.shape
```

```
    (1704, 6)
```

```
df.head(3)
```

```
# A-1: Row oriented
# A-2: Column Oriented
```

```
pd.DataFrame([['Afghanistan',1952, 8425333, 'Asia', 28.801, 779.445314 ],
              ['Afghanistan',1957, 9240934, 'Asia', 30.332, 820.853030 ],
              ['Afghanistan',1962, 102267083, 'Asia', 31.997, 853.100710]],
             columns=['country','year','population','continent','life_exp','gdp_cap'])
```

```
pd.DataFrame([['Afghanistan',1952, 8425333, 'Asia', 28.801, 779.445314]],
             columns=['country','year','population','continent','life_exp','gdp_cap'])
```

```
pd.DataFrame({'country':['Afghanistan', 'Afghanistan'], 'year':[1952,1957],
              'population':[842533, 9240934], 'continent':['Asia', 'Asia'],
              'life_exp':[28.801, 30.332], 'gdp_cap':[779.445314, 820.853030]})
```

```
# Basic Ops on columns
```

```
df.columns
```

```
Index(['country', 'year', 'population', 'continent', 'life_exp', 'gdp_cap'], dtype='object')
```

```
df.keys()
```

```
Index(['country', 'year', 'population', 'continent', 'life_exp', 'gdp_cap'], dtype='object')
```

```
df[["country"]] # now this is a dataframe
```

```
df[["country", "population"]]
```

```
np.unique(df["country"], return_counts=True)
```

```
(array(['Afghanistan', 'Albania', 'Algeria', 'Angola', 'Argentina',
        'Australia', 'Austria', 'Bahrain', 'Bangladesh', 'Belgium',
        'Benin', 'Bolivia', 'Bosnia and Herzegovina', 'Botswana', 'Brazil',
        'Bulgaria', 'Burkina Faso', 'Burundi', 'Cambodia', 'Cameroon',
        'Canada', 'Central African Republic', 'Chad', 'Chile', 'China',
        'Colombia', 'Comoros', 'Congo, Dem. Rep.', 'Congo, Rep.',
        'Costa Rica', "Cote d'Ivoire", 'Croatia', 'Cuba', 'Czech Republic',
        'Denmark', 'Djibouti', 'Dominican Republic', 'Ecuador', 'Egypt',
        'El Salvador', 'Equatorial Guinea', 'Eritrea', 'Ethiopia',
        'Finland', 'France', 'Gabon', 'Gambia', 'Germany', 'Ghana',
        'Greece', 'Guatemala', 'Guinea', 'Guinea-Bissau', 'Haiti',
        'Honduras', 'Hong Kong, China', 'Hungary', 'Iceland', 'India',
        'Indonesia', 'Iran', 'Iraq', 'Ireland', 'Israel', 'Italy',
        'Jamaica', 'Japan', 'Jordan', 'Kenya', 'Korea, Dem. Rep.',
        'Korea, Rep.', 'Kuwait', 'Lebanon', 'Lesotho', 'Liberia', 'Libya',
        'Madagascar', 'Malawi', 'Malaysia', 'Mali', 'Mauritania',
        'Mauritius', 'Mexico', 'Mongolia', 'Montenegro', 'Morocco',
        'Mozambique', 'Myanmar', 'Namibia', 'Nepal', 'Netherlands',
        'New Zealand', 'Nicaragua', 'Niger', 'Nigeria', 'Norway', 'Oman',
        'Pakistan', 'Panama', 'Paraguay', 'Peru', 'Philippines', 'Poland',
        'Portugal', 'Puerto Rico', 'Reunion', 'Romania', 'Rwanda',
        'Sao Tome and Principe', 'Saudi Arabia', 'Senegal', 'Serbia',
        'Sierra Leone', 'Singapore', 'Slovak Republic', 'Slovenia',
        'Somalia', 'South Africa', 'Spain', 'Sri Lanka', 'Sudan',
        'Swaziland', 'Sweden', 'Switzerland', 'Syria', 'Taiwan',
        'Tanzania', 'Thailand', 'Togo', 'Trinidad and Tobago', 'Tunisia',
        'Turkey', 'Uganda', 'United Kingdom', 'United States', 'Uruguay',
        'Venezuela', 'Vietnam', 'West Bank and Gaza', 'Yemen, Rep.',
        'Zambia', 'Zimbabwe'], dtype=object),
 array([12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,
        12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,
        12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,
        12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,
        12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,
        12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,
        12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,
        12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,
        12, 12, 12, 12, 12, 12]))
```

```
df["country"].unique()
```

```
array(['Afghanistan', 'Albania', 'Algeria', 'Angola', 'Argentina',
       'Australia', 'Austria', 'Bahrain', 'Bangladesh', 'Belgium',
       'Benin', 'Bolivia', 'Bosnia and Herzegovina', 'Botswana', 'Brazil',
       'Bulgaria', 'Burkina Faso', 'Burundi', 'Cambodia', 'Cameroon',
       'Canada', 'Central African Republic', 'Chad', 'Chile', 'China',
       'Colombia', 'Comoros', 'Congo, Dem. Rep.', 'Congo, Rep.',
       'Costa Rica', "Cote d'Ivoire", 'Croatia', 'Cuba', 'Czech Republic',
       'Denmark', 'Djibouti', 'Dominican Republic', 'Ecuador', 'Egypt',
       'El Salvador', 'Equatorial Guinea', 'Eritrea', 'Ethiopia',
       'Finland', 'France', 'Gabon', 'Gambia', 'Germany', 'Ghana',
       'Greece', 'Guatemala', 'Guinea', 'Guinea-Bissau', 'Haiti',
       'Honduras', 'Hong Kong, China', 'Hungary', 'Iceland', 'India',
       'Indonesia', 'Iran', 'Iraq', 'Ireland', 'Israel', 'Italy',
       'Jamaica', 'Japan', 'Jordan', 'Kenya', 'Korea, Dem. Rep.',
       'Korea, Rep.', 'Kuwait', 'Lebanon', 'Lesotho', 'Liberia', 'Libya',
       'Madagascar', 'Malawi', 'Malaysia', 'Mali', 'Mauritania',
       'Mauritius', 'Mexico', 'Mongolia', 'Montenegro', 'Morocco',
       'Mozambique', 'Myanmar', 'Namibia', 'Nepal', 'Netherlands',
       'New Zealand', 'Nicaragua', 'Niger', 'Nigeria', 'Norway', 'Oman',
       'Pakistan', 'Panama', 'Paraguay', 'Peru', 'Philippines', 'Poland',
       'Portugal', 'Puerto Rico', 'Reunion', 'Romania', 'Rwanda',
       'Sao Tome and Principe', 'Saudi Arabia', 'Senegal', 'Serbia',
       'Sierra Leone', 'Singapore', 'Slovak Republic', 'Slovenia',
       'Somalia', 'South Africa', 'Spain', 'Sri Lanka', 'Sudan',
       'Swaziland', 'Sweden', 'Switzerland', 'Syria', 'Taiwan',
       'Tanzania', 'Thailand', 'Togo', 'Trinidad and Tobago', 'Tunisia',
       'Turkey', 'Uganda', 'United Kingdom', 'United States', 'Uruguay',
       'Venezuela', 'Vietnam', 'West Bank and Gaza', 'Yemen, Rep.',
       'Zambia', 'Zimbabwe'], dtype=object)
```

```
df["country"].value_counts()
```

```
Afghanistan          12
Pakistan             12
New Zealand          12
Nicaragua            12
Niger                12
                     ..
Eritrea              12
Equatorial Guinea    12
El Salvador          12
Egypt                12
Zimbabwe             12
Name: country, Length: 142, dtype: int64
```

```
df.rename({"population": "Population", "country":"Country" }, axis = 1)
df.rename(columns={"country":"Country"}) # wont suggest
```

```
df.rename({"country": "Country"}, axis = 1, inplace = True)
df
```

```
df
```

```
df["Country"]
```

```
0       Afghanistan
1       Afghanistan
2       Afghanistan
3       Afghanistan
4       Afghanistan
           ...
1699       Zimbabwe
1700       Zimbabwe
1701       Zimbabwe
1702       Zimbabwe
1703       Zimbabwe
Name: Country, Length: 1704, dtype: object
```

```
df.Country # SERIOUSLY not recommended
# homework
# column name --> shape, shape is also an attribute to extract the shape
# roll number df.roll number
```

```
0       Afghanistan
1       Afghanistan
2       Afghanistan
3       Afghanistan
4       Afghanistan
           ...
1699       Zimbabwe
1700       Zimbabwe
1701       Zimbabwe
1702       Zimbabwe
1703       Zimbabwe
Name: Country, Length: 1704, dtype: object
```

```
df.drop("continent", axis=1)
```

```
df["year+7"] = df["year"] + 7
```

```
df
```

```
df["gdp"] = df["gdp_cap"] * df["population"]
```

```
df
```

```
df["Own"] = [i for i in range(1704)]
df
```

```
df.drop(["Own", "gdp", "year+7"], axis=1, inplace=True)
```

```
df
```

Pandas1b

```
df
```

```
df.index.values # explcit indices
```

```
array([   0,    1,    2, ..., 1701, 1702, 1703])
```

```
df.index = list(range(1, df.shape[0]+1))
# df.columns
```

```
df
```

```
df.index[1]
```

```
2
```

```
df.index = np.arange(1, df.shape[0]+1, dtype="float")
```

```
df
```

```
sample = df.head()
sample
```

```
sample.index = ["a", "b", "c", "d", "e"]
sample
```

```
df.index = np.arange(1, df.shape[0]+1, dtype='int')
```

```
df
```

```
ser = df["Country"]
ser
```

```
1       Afghanistan
2       Afghanistan
3       Afghanistan
4       Afghanistan
5       Afghanistan
           ...
1700       Zimbabwe
1701       Zimbabwe
```

```
    1702       Zimbabwe
    1703       Zimbabwe
    1704       Zimbabwe
    Name: Country, Length: 1704, dtype: object
```

```
ser[12] # explicit index
```

```
ser[0]
```

```
ser[4:15] # impliit index
```

```
    5     Afghanistan
    6     Afghanistan
    7     Afghanistan
    8     Afghanistan
    9     Afghanistan
    10    Afghanistan
    11    Afghanistan
    12    Afghanistan
    13        Albania
    14        Albania
    15        Albania
    Name: Country, dtype: object
```

```
# this is the case with series
# indexing --> explicit index
# slicing --> implcit index
```

```
df[1] # df["country"] ---> using explcit index
# this syntax is used for accessing columns
# not for rows
```

```
df[4:15]
```

```
# Series
# indexing --> explicit index
# slicing --> implcit index

# Dataframe
# indexing --> doesn't work, used for columns df["country"], df[1]
# slicing --> implicit index
```

Indexers (loc and iloc)

```
df
```

```
df.loc[1]
```

```
    Country       Afghanistan
    year                 1952
    population        8425333
    continent            Asia
    life_exp           28.801
    gdp_cap        779.445314
    Name: 1, dtype: object
```

```
df.loc[1:3] # end point in explcit indexing is included
```

```
df.iloc[1]
```

```
    Country       Afghanistan
    year                 1957
    population        9240934
    continent            Asia
    life_exp           30.332
    gdp_cap         820.85303
    Name: 2, dtype: object
```

```
df.iloc[0:2] # implcit indexing doesnt include end point
```

```
df.iloc[[2, 10, 100]]
```

```
df.iloc[-1]
```

```
    Country         Zimbabwe
    year                2007
    population      12311143
    continent         Africa
    life_exp          43.487
    gdp_cap       469.709298
    Name: 1704, dtype: object
```

```
df.loc[-1]
```

```
temp = df.set_index("Country")
temp
```

```
temp.loc["Afghanistan"]
```

```
temp.reset_index()
```

```
df
```

```
df.reset_index(drop=True, inplace=True)
```

```
df
```

```
# append
# loc/iloc
```

```
new_row = {'Country': 'India',
           'year': 2000,
           'life_exp':37.08,
           'population':13500000,
           'continent': "Asia",
           'gdp_cap':900.23}
```

```
df.append(new_row, ignore_index=True)
df.loc[1704] = ['India',2000 ,13500000, "Asia", 37.08,900.23]
df.loc[1705] = ['India',2000 ,13500000, "Asia", 37.08,900.23]
```

```
df
```

```
df.iloc[1705] = ['India',2000 ,13500000, "Asia", 37.08,900.23]
```

```
df.iloc[1706] = ['India',2000 ,13500000, "Asia", 37.08,900.23]
```

```
df.drop([1, 2, 4], axis=0)
```

```
df.loc[len(df.index)] = ['India',2000,13500000,"Asia", 37.08,900.23]
df.loc[len(df.index)] = ['Sri Lanka',2022 ,130000000,"Asia", 80.00,500.00]
df.loc[len(df.index)] = ['Sri Lanka',2022 ,130000000,"Asia",80.00,500.00]
df.loc[len(df.index)] = ['India',2000 ,13500000,"Asia",80.00,900.23]
df
```

```
df.duplicated()
```

```
0       False
1       False
2       False
3       False
4       False
        ...
1705     True
1706     True
1707    False
1708     True
1709    False
Length: 1710, dtype: bool
```

```
df.loc[df.duplicated(), :]
```

```
df.drop_duplicates(keep="last") # keep = False, remove all
```

● ×