Colab: https://colab.research.google.com/drive/15Gh2h7I2erNPqfXd7e-Ok8pldEG9kf0N?usp=sharing

```
import pandas as pd
```

+ Code + Text

```
import numpy as np


# GDP per capita V/S Life Expectancy


!wget "https://drive.google.com/uc?export=download&id=1E3bwvYGf1ig32RmcYiWc0IXPN-mD
```

```
--2022-12-05 15:30:18--  https://drive.google.com/uc?export=download&id=1E3bwv
Resolving drive.google.com (drive.google.com)... 142.250.141.139, 142.250.141.
Connecting to drive.google.com (drive.google.com)|142.250.141.139|:443... conn
HTTP request sent, awaiting response... 303 See Other
Location: https://doc-0s-68-docs.googleusercontent.com/docs/securesc/ha0ro937c
Warning: wildcards not supported in HTTP.
--2022-12-05 15:30:19--  https://doc-0s-68-docs.googleusercontent.com/docs/sec
Resolving doc-0s-68-docs.googleusercontent.com (doc-0s-68-docs.googleuserconte
Connecting to doc-0s-68-docs.googleusercontent.com (doc-0s-68-docs.googleuserc
HTTP request sent, awaiting response... 200 OK
Length: 83785 (82K) [text/csv]
Saving to: 'gapminder.csv'

gapminder.csv       100%[===================>]  81.82K  --.-KB/s    in 0.001s

2022-12-05 15:30:19 (93.3 MB/s) - 'gapminder.csv' saved [83785/83785]
```

```
df = pd.read_csv("gapminder.csv")
df # structured data, tabular data
```

	country	year	population	continent	life_exp	gdp_cap
0	Afghanistan	1952	8425333	Asia	28.801	779.445314

```
type(df)
```

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

3	Afghanistan	1967	11537966	Asia	34.020	836.197138

```
df.shape
```

```
(1704, 6)
```

```
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
```

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

```
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)
```

	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	Asia	34.020	836.197138

```
df.tail(6)
```

	country	year	population	continent	life_exp	gdp_cap
1698	Zimbabwe	1982	7636524	Africa	60.363	788.855041
1699	Zimbabwe	1987	9216418	Africa	62.351	706.157306
1700	Zimbabwe	1992	10704340	Africa	60.377	693.420786
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

```
# row orineted approach - lists of lists

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_c
```

	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	102267083	Asia	31.997	853.100710

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

	country	year	population	continent	life_exp	gdp_cap
0	Afghanistan	1952	8425333	Asia	28.801	779.445314

```
# column oriented approach

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]})
```

	country	year	population	continent	life_exp	gdp_cap
0	Afghanistan	1952	842533	Asia	28.801	779.445314
1	Afghanistan	1957	9240934	Asia	30.332	820.853030

```
# learner's doubt - columns with same names, gets replaced
pd.DataFrame({'country':['Afghanistan', 'Afghanistan'], 'year':[1952,1957],
              'population':[842533, 9240934], 'continent':['Asia', 'Asia'],
              'life_exp':[28.801, 30.332], 'continent':[779.445314, 820.853030]})
```

	country	year	population	continent	life_exp
0	Afghanistan	1952	842533	779.445314	28.801
1	Afghanistan	1957	9240934	820.853030	30.332

```
df.columns
```

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

```
df.keys() # df works like a "specialised" dictionary
```

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

```
df[["country", "life_exp"]] # extract subset of columns
```

	country	life_exp
0	Afghanistan	28.801
1	Afghanistan	30.332
2	Afghanistan	31.997
3	Afghanistan	34.020
4	Afghanistan	36.088
...	...	...
1699	Zimbabwe	62.351
1700	Zimbabwe	60.377
1701	Zimbabwe	46.809
1702	Zimbabwe	39.989
1703	Zimbabwe	43.487

1704 rows × 2 columns

```
df["country"] # series
```

```
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"]] # returns as dataframe because of double brackers
```

	country
0	Afghanistan
1	Afghanistan
2	Afghanistan
3	Afghanistan
4	Afghanistan
...	...
1699	Zimbabwe
1700	Zimbabwe
1701	Zimbabwe
1702	Zimbabwe
1703	Zimbabwe

1704 rows × 1 columns

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

```
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"].nunique()
```

```
142
```

```
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, inplace=True)
```

```
df
```

	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

```
df
```

	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	Asia	34.020	836.197138
4	Afghanistan	1972	13079460	Asia	36.088	739.981106
...	...	...	...	...	...	...
1699	Zimbabwe	1987	9216418	Africa	62.351	706.157306
1700	Zimbabwe	1992	10704340	Africa	60.377	693.420786
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

1704 rows × 6 columns

```
df.rename({"continent":"Continent"}) # default axis=0
```

	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	Asia	34.020	836.197138
4	Afghanistan	1972	13079460	Asia	36.088	739.981106
...	...	...	...	...	...	...
1699	Zimbabwe	1987	9216418	Africa	62.351	706.157306
1700	Zimbabwe	1992	10704340	Africa	60.377	693.420786
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

1704 rows × 6 columns

```
df["Country"] # dict like style
```

```
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 # attribute, DONT USE THIS
```

```
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
```

```
# HOMEWORK - WHY I SHOULD NOT USE THIS STYLE
```

```
df.drop("continent", axis=1) # for permanent changes use inplace=True
```

	Country	year	Population	life_exp	gdp_cap

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

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

3	Afghanistan	1967	11537966	34.020	836.197138

```
df
```

	Country	year	Population	continent	life_exp	gdp_cap	year+7	
0	Afghanistan	1952	8425333	Asia	28.801	779.445314	1959	6.567(
1	Afghanistan	1957	9240934	Asia	30.332	820.853030	1964	7.585-
2	Afghanistan	1962	10267083	Asia	31.997	853.100710	1969	8.758
3	Afghanistan	1967	11537966	Asia	34.020	836.197138	1974	9.648(
4	Afghanistan	1972	13079460	Asia	36.088	739.981106	1979	9.678
...	...	...	...	...	...	...	...	
1699	Zimbabwe	1987	9216418	Africa	62.351	706.157306	1994	6.508
1700	Zimbabwe	1992	10704340	Africa	60.377	693.420786	1999	7.422(
1701	Zimbabwe	1997	11404948	Africa	46.809	792.449960	2004	9.037
1702	Zimbabwe	2002	11926563	Africa	39.989	672.038623	2009	8.015
1703	Zimbabwe	2007	12311143	Africa	43.487	469.709298	2014	5.782(

1704 rows × 8 columns

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

```
df
```

	Country	year	Population	continent	life_exp	gdp_cap	year+7	
0	Afghanistan	1952	8425333	Asia	28.801	779.445314	1959	6.567(
1	Afghanistan	1957	9240934	Asia	30.332	820.853030	1964	7.585⁴

```
# df.insert() that helps in adding a column at a particular location
```

3	Afghanistan	1967	11537966	Asia	34.020	836.197138	1974	9.648(

```
# Working with Rows


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

```
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
```

```
ser[0]
```

```
'Afghanistan'
```

```
ser[5:14]
```

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

```
df[0] # indexing a row like doesnt happan in dataframe
# df["country"] #looks for this index along axis=1
```

```
---------------------------------------------------------------------------
KeyError                                  Traceback (most recent call last)
/usr/local/lib/python3.8/dist-packages/pandas/core/indexes/base.py in
get_loc(self, key, method, tolerance)
   3360             try:
-> 3361                 return self._engine.get_loc(casted_key)
   3362             except KeyError as err:

                                4 frames

pandas/_libs/hashtable_class_helper.pxi in
pandas._libs.hashtable.PyObjectHashTable.get_item()

pandas/_libs/hashtable_class_helper.pxi in
pandas._libs.hashtable.PyObjectHashTable.get_item()

KeyError: 0

The above exception was the direct cause of the following exception:

KeyError                                  Traceback (most recent call last)
/usr/local/lib/python3.8/dist-packages/pandas/core/indexes/base.py in
get_loc(self, key, method, tolerance)
   3361                 return self._engine.get_loc(casted_key)
```

```
df[5:14]
```

	Country	year	Population	continent	life_exp	gdp_cap	year+7	
5	Afghanistan	1977	14880372	Asia	38.438	786.113360	1984	1.16976
6	Afghanistan	1982	12881816	Asia	39.854	978.011439	1989	1.25985
7	Afghanistan	1987	13867957	Asia	40.822	852.395945	1994	1.18209
8	Afghanistan	1992	16317921	Asia	41.674	649.341395	1999	1.05959
9	Afghanistan	1997	22227415	Asia	41.763	635.341351	2004	1.41220
10	Afghanistan	2002	25268405	Asia	42.129	726.734055	2009	1.83634
11	Afghanistan	2007	31889923	Asia	43.828	974.580338	2014	3.10792
12	Albania	1952	1282697	Europe	55.230	1601.056136	1959	2.05367
13	Albania	1957	1476505	Europe	59.280	1942.284244	1964	2.86779

```
# ===> Indexing doesnt work to index rows because of similar syntax for columns
# ===> Slicing works for slicing the rows
```

```
df.index.values
```

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

```
df.columns
```

```
Index(['Country', 'year', 'Population', 'continent', 'life_exp', 'gdp_cap',
       'year+7', 'gdp', 'Own'],
      dtype='object')
```

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

```
df
```

	Country	year	Population	continent	life_exp	gdp_cap	year+7	
1	Afghanistan	1952	8425333	Asia	28.801	779.445314	1959	6.567(
2	Afghanistan	1957	9240934	Asia	30.332	820.853030	1964	7.585⁴
3	Afghanistan	1962	10267083	Asia	31.997	853.100710	1969	8.758ᵹ
4	Afghanistan	1967	11537966	Asia	34.020	836.197138	1974	9.648(
5	Afghanistan	1972	13079460	Asia	36.088	739.981106	1979	9.678ᵹ
...	...	...	...	...	...	...	...	
1700	Zimbabwe	1987	9216418	Africa	62.351	706.157306	1994	6.508ᵹ
1701	Zimbabwe	1992	10704340	Africa	60.377	693.420786	1999	7.422(
1702	Zimbabwe	1997	11404948	Africa	46.809	792.449960	2004	9.037ᵹ
1703	Zimbabwe	2002	11926563	Africa	39.989	672.038623	2009	8.015'
1704	Zimbabwe	2007	12311143	Africa	43.487	469.709298	2014	5.782(

1704 rows × 9 columns

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

```
2
```

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

	Country	year	Population	continent	life_exp	gdp_cap	year+7	
1.0	Afghanistan	1952	8425333	Asia	28.801	779.445314	1959	6.56

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

```
sample
```

	Country	year	Population	continent	life_exp	gdp_cap	year+7	
1.0	Afghanistan	1952	8425333	Asia	28.801	779.445314	1959	6.56708
2.0	Afghanistan	1957	9240934	Asia	30.332	820.853030	1964	7.58544
3.0	Afghanistan	1962	10267083	Asia	31.997	853.100710	1969	8.75885
4.0	Afghanistan	1967	11537966	Asia	34.020	836.197138	1974	9.64801
5.0	Afghanistan	1972	13079460	Asia	36.088	739.981106	1979	9.67855

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

```
sample
```

	Country	year	Population	continent	life_exp	gdp_cap	year+7	
a	Afghanistan	1952	8425333	Asia	28.801	779.445314	1959	6.567086e
b	Afghanistan	1957	9240934	Asia	30.332	820.853030	1964	7.585449e
c	Afghanistan	1962	10267083	Asia	31.997	853.100710	1969	8.758856e
d	Afghanistan	1967	11537966	Asia	34.020	836.197138	1974	9.648014e
e	Afghanistan	1972	13079460	Asia	36.088	739.981106	1979	9.678553e

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

```
sample
```

	Country	year	Population	continent	life_exp	gdp_cap	year+7	
a	Afghanistan	1952	8425333	Asia	28.801	779.445314	1959	6.567086e
b	Afghanistan	1957	9240934	Asia	30.332	820.853030	1964	7.585449e
c	Afghanistan	1962	10267083	Asia	31.997	853.100710	1969	8.758856e
d	Afghanistan	1967	11537966	Asia	34.020	836.197138	1974	9.648014e
d	Afghanistan	1972	13079460	Asia	36.088	739.981106	1979	9.678553e

```
df.columns
```

```
Index(['Country', 'year', 'Population', 'continent', 'life_exp', 'gdp_cap',
       'year+7', 'gdp', 'Own'],
      dtype='object')
```

```
df.columns = ['Country', 'year', 'Population', 'continent', 'life_exp', 'gdp_cap',
       'gdp', 'Country']
```

```
---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
<ipython-input-60-9856190f8687> in <module>
----> 1 df.columns = ['Country', 'year', 'Population', 'continent', 'life_exp'
'gdp_cap',
      2        'gdp', 'Country']

                                4 frames
/usr/local/lib/python3.8/dist-packages/pandas/core/internals/base.py in
_validate_set_axis(self, axis, new_labels)
     55
     56         elif new_len != old_len:
---> 57             raise ValueError(
     58                 f"Length mismatch: Expected axis has {old_len} element
new "
     59                 f"values have {new_len} elements"

ValueError: Length mismatch: Expected axis has 9 elements, new values have 8
```

```
df
```

```
# duplication of explciit indexes is ok
```

```
df["Country"] # you can basically group the data, classification using same explcit
```

```
1.0       Afghanistan
2.0       Afghanistan
3.0       Afghanistan
4.0       Afghanistan
5.0       Afghanistan
             ...
1700.0       Zimbabwe
1701.0       Zimbabwe
1702.0       Zimbabwe
1703.0       Zimbabwe
1704.0       Zimbabwe
Name: Country, Length: 1704, dtype: object
```

```
df.reset_index() # drop=True would have dropped the colum
```

	index	Country	year	Population	continent	life_exp	gdp_cap	year+7
0	1.0	Afghanistan	1952	8425333	Asia	28.801	779.445314	1959
1	2.0	Afghanistan	1957	9240934	Asia	30.332	820.853030	1964
2	3.0	Afghanistan	1962	10267083	Asia	31.997	853.100710	1969
3	4.0	Afghanistan	1967	11537966	Asia	34.020	836.197138	1974
4	5.0	Afghanistan	1972	13079460	Asia	36.088	739.981106	1979
...	...	...	...	...	...	...	...	..
1699	1700.0	Zimbabwe	1987	9216418	Africa	62.351	706.157306	1994
1700	1701.0	Zimbabwe	1992	10704340	Africa	60.377	693.420786	1999
1701	1702.0	Zimbabwe	1997	11404948	Africa	46.809	792.449960	2004

```
# Learner's personal doubt
```

1703	1704.0	Zimbabwe	2007	12311143	Africa	43.487	469.709298	2014

Double-click (or enter) to edit

```
x = pd.DataFrame({'veclocity':[100, -11, -16, 13, 14]})


def find_closest(x):
  return np.argmin(np.abs(np.array([3, 30, -5, -15]) - x))


x["veclocity"].apply(find_closest)
```

```
0    1
1    3
2    3
3    0
4    0
Name: veclocity, dtype: int64
```

New content

[ ] ↳ *79 cells hidden*