Pandas DataFrame exercises

```
# Import the numpy package under the name np
import numpy as np

# Import the pandas package under the name pd
import pandas as pd

# Import the matplotlib package under the name plt
import matplotlib.pyplot as plt
%matplotlib inline

# Print the pandas version and the configuration
print(pd.__version__)
```

```
1.5.3
```

DataFrame creation

Create an empty pandas DataFrame

```
# your code goes here
pd.DataFrame(data=[None],
             index=[None],
             columns=[None])
```

	None
None	None

Create a `marvel_df` pandas DataFrame with the given marvel data

```
marvel_data = [
    ['Spider-Man', 'male', 1962],
    ['Captain America', 'male', 1941],
    ['Wolverine', 'male', 1974],
    ['Iron Man', 'male', 1963],
    ['Thor', 'male', 1963],
    ['Thing', 'male', 1961],
    ['Mister Fantastic', 'male', 1961],
    ['Hulk', 'male', 1962],
    ['Beast', 'male', 1963],
    ['Invisible Woman', 'female', 1961],
    ['Storm', 'female', 1975],
    ['Namor', 'male', 1939],
    ['Hawkeye', 'male', 1964],
    ['Daredevil', 'male', 1964],
    ['Doctor Strange', 'male', 1963],
    ['Hank Pym', 'male', 1962],
    ['Scarlet Witch', 'female', 1964],
    ['Wasp', 'female', 1963],
    ['Black Widow', 'female', 1964],
    ['Vision', 'male', 1968]
]
```

```
# your code goes here
marvel = pd.DataFrame(data=marvel_data)
marvel
```

	0	1	2
0	Spider-Man	male	1962
1	Captain America	male	1941
2	Wolverine	male	1974
3	Iron Man	male	1963
4	Thor	male	1963
5	Thing	male	1961
6	Mister Fantastic	male	1961
7	Hulk	male	1962
8	Beast	male	1963
9	Invisible Woman	female	1961
10	Storm	female	1975
11	Namor	male	1939
12	Hawkeye	male	1964
13	Daredevil	male	1964
14	Doctor Strange	male	1963
15	Hank Pym	male	1962
16	Scarlet Witch	female	1964
17	Wasp	female	1963
18	Black Widow	female	1964
19	Vision	male	1968

Add column names to the `marvel_df`

```
# your code goes here
columns_name = ['name', 'sex', 'first_apperances']

marvel.columns = columns_name
marvel
```

	name	sex	first_apperances
0	Spider-Man	male	1962
1	Captain America	male	1941
2	Wolverine	male	1974
3	Iron Man	male	1963
4	Thor	male	1963
5	Thing	male	1961
6	Mister Fantastic	male	1961
7	Hulk	male	1962
8	Beast	male	1963
9	Invisible Woman	female	1961
10	Storm	female	1975
11	Namor	male	1939
12	Hawkeye	male	1964
13	Daredevil	male	1964
14	Doctor Strange	male	1963
15	Hank Pym	male	1962
16	Scarlet Witch	female	1964
17	Wasp	female	1963
18	Black Widow	female	1964

Add index names to the `marvel_df` (use the character name as index)

```
# your code goes here
marvel.index = marvel['name']
marvel
```

name sex first_apperances

name

Drop the name column as it's now the index

```
# your code goes here
#marvel = marvel.drop(columns=['name'])
marvel = marvel.drop(['name'], axis=1)
marvel
```

name	sex	first_apperances
Spider-Man	male	1962
Captain America	male	1941
Wolverine	male	1974
Iron Man	male	1963
Thor	male	1963
Thing	male	1961
Mister Fantastic	male	1961
Hulk	male	1962
Beast	male	1963
Invisible Woman	female	1961
Storm	female	1975
Namor	male	1939
Hawkeye	male	1964
Daredevil	male	1964
Doctor Strange	male	1963
Hank Pym	male	1962
Scarlet Witch	female	1964
Wasp	female	1963
Black Widow	female	1964
Vision	male	1968

Drop 'Namor' and 'Hank Pym' rows

```
# your code goes here
marvel = marvel.drop(['Namor', 'Hank Pym'], axis=0)
marvel
```

name	sex	first_apperances
Spider-Man	male	1962
Captain America	male	1941
Wolverine	male	1974
Iron Man	male	1963
Thor	male	1963
Thing	male	1961
Mister Fantastic	male	1961
Hulk	male	1962
Beast	male	1963

DataFrame selection, slicing and indexation

Daredevil	male	1964

Show the first 5 elements on `marvel_df`

Scarlet Witch	female	1964

```
# your code goes here
marvel.head()
#marvel.loc[['Spider-Man', 'Captain America', 'Wolverine', 'Iron Man', 'Thor'], :] # bad!
#marvel.loc['Spider-Man': 'Thor', :]
#marvel.iloc[0:5, :]
#marvel.iloc[0:5,]
#marvel.iloc[:5,]
```

name	sex	first_apperances
Spider-Man	male	1962
Captain America	male	1941
Wolverine	male	1974
Iron Man	male	1963
Thor	male	1963

Show the last 5 elements on `marvel_df`

```
# your code goes here
#marvel.loc[['Hank Pym', 'Scarlet Witch', 'Wasp', 'Black Widow', 'Vision'], :] # bad!
#marvel.loc['Hank Pym':'Vision', :]
#marvel.iloc[-5:,]
marvel.tail()
```

name	sex	first_apperances
Doctor Strange	male	1963
Scarlet Witch	female	1964
Wasp	female	1963
Black Widow	female	1964
Vision	male	1968

Show just the sex of the first 5 elements on `marvel_df`

```
# your code goes here
#marvel.iloc[:5,]['sex'].to_frame()
marvel.iloc[:5,].sex.to_frame()
#marvel.head().sex.to_frame()
```

name	sex
Spider-Man	male
Captain America	male
Wolverine	male
Iron Man	male
Thor	male

Show the first_appearance of all middle elements on `marvel_df`

```
# your code goes here
marvel.iloc[1:-1,].first_apperances.to_frame()
#mistake in adding column name
```

name	first_apperances
Captain America	1941
Wolverine	1974
Iron Man	1963
Thor	1963
Thing	1961
Mister Fantastic	1961
Hulk	1962
Beast	1963
Invisible Woman	1961
Storm	1975
Hawkeye	1964
Daredevil	1964
Doctor Strange	1963
Scarlet Witch	1964
Wasp	1963
Black Widow	1964

Show the first and last elements on `marvel_df`

```
# your code goes here
#marvel_df.iloc[[0, -1],][['sex', 'first_apperances']]
marvel.iloc[[0, -1],]
```

sex first_apperances

DataFrame manipulation and operations

Modify the `first_appearance` of 'Vision' to year 1964

```
# your code goes here
marvel.loc['Vision', 'first_apperances'] = 1964
#marvel.loc['Index', 'Column'] = new_value
marvel
```

name	sex	first_apperances
Spider-Man	male	1962
Captain America	male	1941
Wolverine	male	1974
Iron Man	male	1963
Thor	male	1963
Thing	male	1961
Mister Fantastic	male	1961
Hulk	male	1962
Beast	male	1963
Invisible Woman	female	1961
Storm	female	1975
Hawkeye	male	1964
Daredevil	male	1964
Doctor Strange	male	1963
Scarlet Witch	female	1964
Wasp	female	1963
Black Widow	female	1964
Vision	male	1964

Add a new column to `marvel_df` called 'years_since' with the years since `first_appearance`

```
# your code goes here
marvel['years_sice'] = 2023 - marvel['first_apperances']
marvel
```

name	sex	first_apperances	years_sice
Spider-Man	male	1962	61
Captain America	male	1941	82
Wolverine	male	1974	49
Iron Man	male	1963	60
Thor	male	1963	60
Thing	male	1961	62
Mister Fantastic	male	1961	62
Hulk	male	1962	61
Hawkeye	male	1964	59

DataFrame boolean arrays (also called masks)

Given the `marvel_df` pandas DataFrame, make a mask showing the female characters

```
# your code goes here
mask = marvel['sex'] == 'female'
mask
```

```
name
Spider-Man          False
Captain America     False
Wolverine           False
Iron Man            False
Thor                False
Thing               False
Mister Fantastic    False
Hulk                False
Beast               False
Invisible Woman      True
Storm                True
Hawkeye             False
Daredevil           False
Doctor Strange      False
Scarlet Witch        True
Wasp                 True
Black Widow          True
Vision              False
Name: sex, dtype: bool
```

Given the `marvel_df` pandas DataFrame, get the male characters

```
# your code goes here
mask = marvel['sex'] == 'male'

marvel[mask]
```

name	sex	first_apperances	years_sice
Spider-Man	male	1962	61
Captain America	male	1941	82
Wolverine	male	1974	49
Iron Man	male	1963	60

Given the `marvel_df` pandas DataFrame, get the characters with `first_appearance` after 1970

```
# your code goes here
mask = marvel['first_apperances'] > 1970

marvel[mask]
```

name	sex	first_apperances	years_sice
Wolverine	male	1974	49
Storm	female	1975	48

Given the `marvel_df` pandas DataFrame, get the female characters with `first_appearance` after 1970

```
# your code goes here
mask = (marvel ['sex'] == 'female') & (marvel['first_apperances'] > 1970)

marvel[mask]
```

name	sex	first_apperances	years_sice
Storm	female	1975	48

DataFrame summary statistics

Show basic statistics of `marvel_df`

```
# your code goes here
marvel.describe()
```

	first_apperances	years_sice
count	18.000000	18.000000
mean	1962.888889	60.111111
std	6.720372	6.720372
min	1941.000000	48.000000
25%	1962.000000	59.000000
50%	1963.000000	60.000000
75%	1964.000000	61.000000
max	1975.000000	82.000000

```
# your code goes here
marvel.first_apperances.mean()
```

```
1962.888888888889
```

- Given the `marvel_df` pandas DataFrame, show the min value of `first_appearance`

```
# your code goes here
marvel.first_apperances.min()
```

```
1941
```

- Given the `marvel_df` pandas DataFrame, get the characters with the min value of `first_appearance`

```
# your code goes here
mask = marvel['first_apperances'] == marvel.first_apperances.min()
marvel[mask]
```

name	sex	first_apperances	years_sice
Captain America	male	1941	82

- DataFrame basic plottings

- Reset index names of `marvel_df`

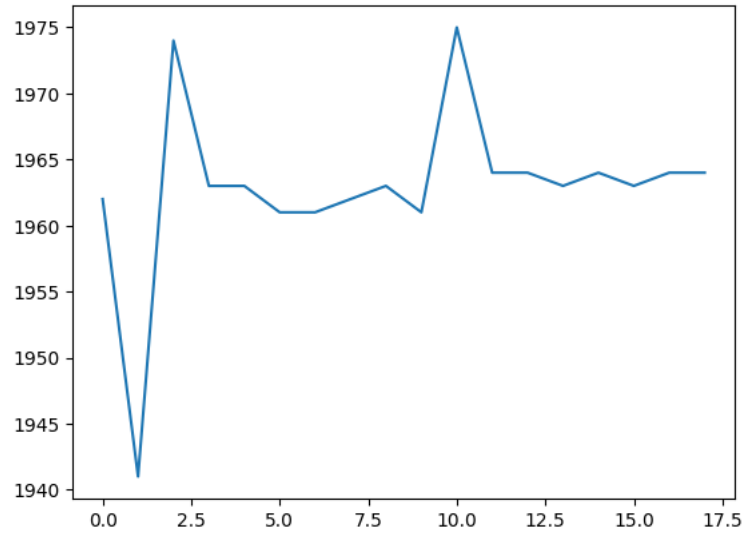
```
# your code goes here
marvel = marvel.reset_index()
marvel
```

	name	sex	first_apperances	years_sice
0	Spider-Man	male	1962	61
1	Captain America	male	1941	82
2	Wolverine	male	1974	49
3	Iron Man	male	1963	60
4	Thor	male	1963	60
5	Thing	male	1961	62
6	Mister Fantastic	male	1961	62
7	Hulk	male	1962	61
8	Beast	male	1963	60
9	Invisible Woman	female	1961	62
10	Storm	female	1975	48
11	Hawkeye	male	1964	59
12	Daredevil	male	1964	59
13	Doctor Strange	male	1963	60
14	Scarlet Witch	female	1964	59
15	Wasp	female	1963	60
16	Black Widow	female	1964	59
17	Vision	male	1964	59

Plot the values of `first_appearance`

```
# your code goes here
marvel.first_apperances.plot()
```

```
<Axes: >
```



Plot a histogram (plot.hist) with values of `first_appearance`

```
# your code goes here
plt.hist(marvel.first_apperances)
```

```
(array([ 1.,  0.,  0.,  0.,  0.,  3., 12.,  0.,  0.,  2.]),
 array([1941. , 1944.4, 1947.8, 1951.2, 1954.6, 1958. , 1961.4, 1964.8,
        1968.2, 1971.6, 1975. ]),
 <BarContainer object of 10 artists>)
```