Pandas1-Objects

Course content

Part 1: Basics of programming (W34-40)

- Programming languages
- Programming environment
- Python syntax
- Variables
- Strings and numbers
- Lists
- Conditional statements
- Loop statements
- Functions
- Dictionary, tuple, and set

Part 2: Data extraction and visualization (W41-47)

- Pandas Series and DataFrame
- Read csv files
- Basic statistics
- Data manipulation
- Missing data handling
- Data aggregation
- Data visualization – Pandas
- Data visualization - Matplotlib
- Data visualization - Seaborn
- Time series data

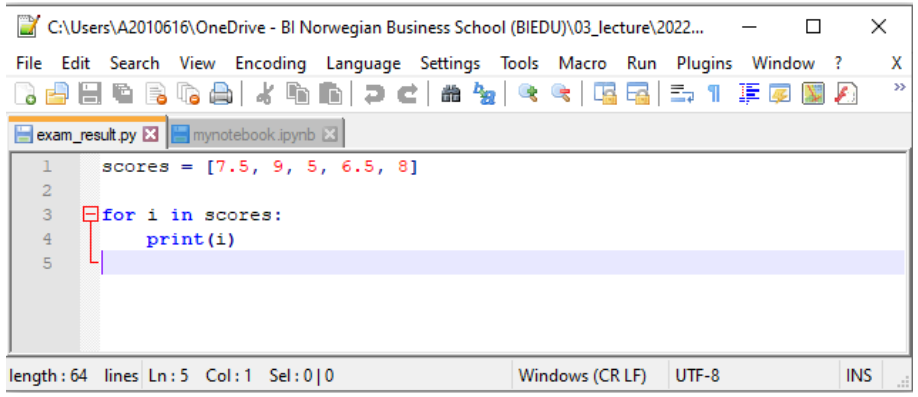
Outline

- **What is a Python package**
  - .py file
  - Modules
  - Packages
  - User-defined class
- **Pandas package**
  - Pandas data structure
    - Series: Creation, index, subset selection, filter
    - DataFrame: Creation, index, subset selection, filter
  - Read file as a DataFrame
    - View data
    - Subset selection
    - Descriptive statistics

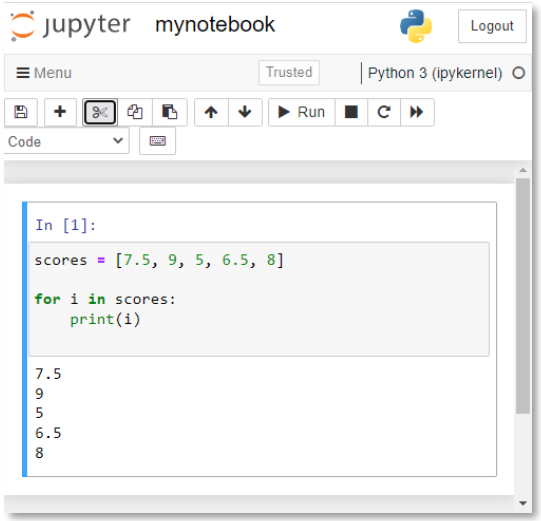
Python file

- .py is a regular python file. It's plain text and contains just your code.
- .ipynb is an interactive python notebook and it contains the notebook code, the execution results and other internal settings in a specific format.

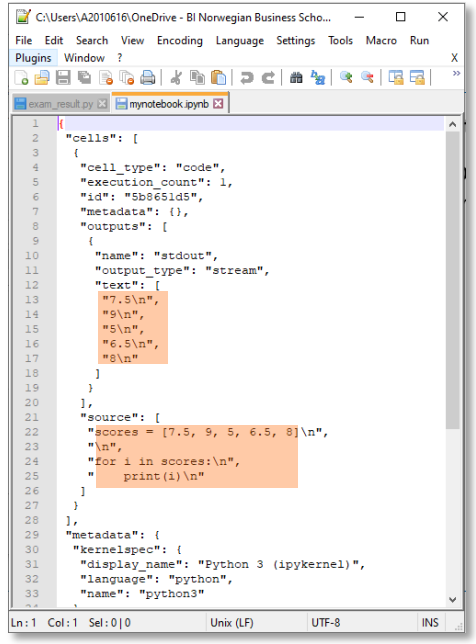
Open .py file in Notepad



.ipynb file



Open .ipynb file in Notepad

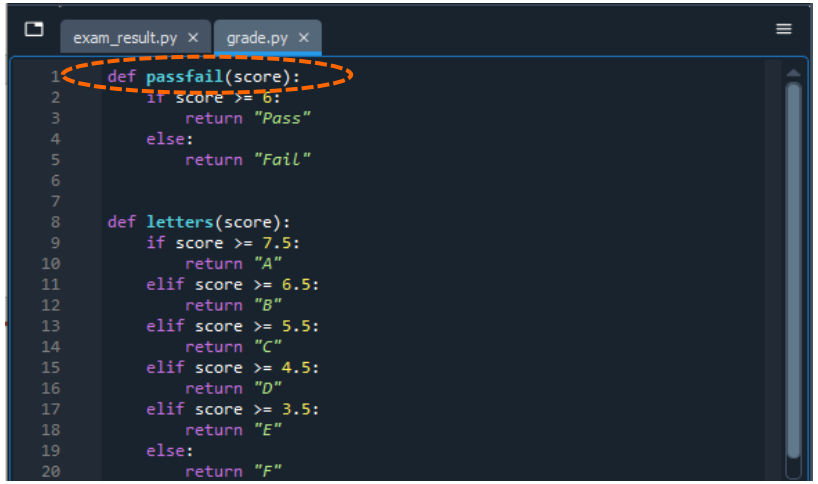


- For mac users, try to use the app TextEdit to open the file.
- The Jupyter Notebook format follows JSON schema. https://nbformat.readthedocs.io/en/latest/format_description.html

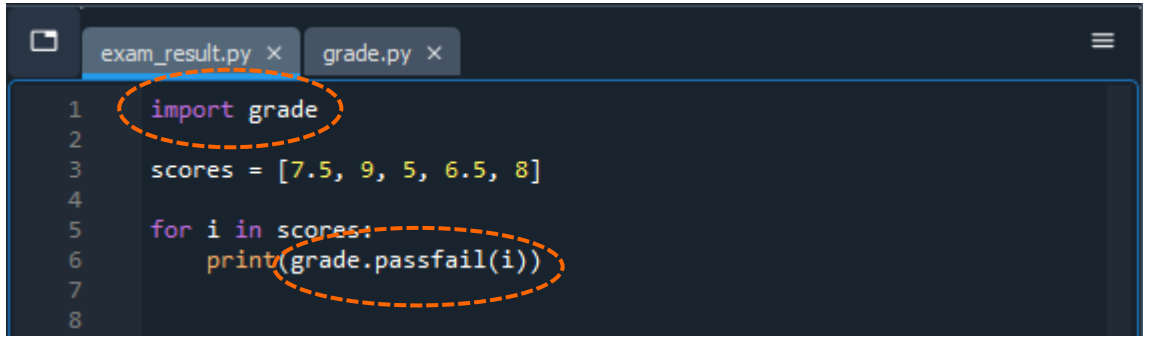
Modules

- Modules:
  - A module is a .py file with python code. The code can be in the form of variables, functions, or class defined. The filename is the module name.
  - The function or variables present inside the file can be used in another file. To use the module, you can import it using the `import` keyword.

Module: grade.py



Script: exam_result.py



- https://docs.python.org/3/tutorial/modules.html#

Optional exercise: Create your own module

- Step1: Create a module named **student**. (Use **student.py** as file name).
- Step2: Write some functions and save the file.
- Step3: Under the same folder, create a jupyter notebook.
- Step4: Import the module **“student”.**
- Step5: Call the function **greeting()** from the module.

student.py

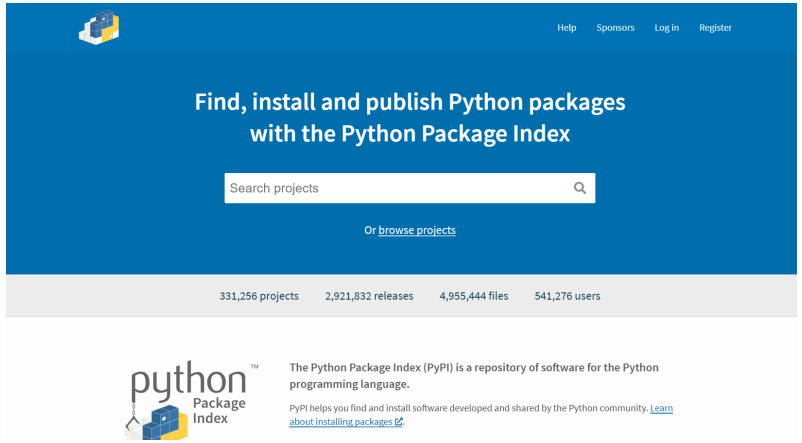
```python
def greeting():
    print("Welcome to BI!")

def profile(student_id, name):
    print("New Student Profile:")
    print(f"ID: {student_id}")
    print(f"Name: {name}")

```

```
In [1]: import student

In [2]: student.greeting()
        Welcome to BI!

In [3]: student.profile("S01234", "Anna")
        New Student Profile:
        ID: S01234
        Name: Anna
```

Packages

- Packages are a collection of modules.
- The **Py**thon **P**ackage **I**ndex (PyPI) is a repository of pacakges for the Python programming language.
- **pip** is the package installer for Python. You can use it to install packages from PyPI.

https://pypi.org/



https://docs.python.org/3/tutorial/modules.html#packages

Packages – packages for data analysis

- **NumPy** (Numerical Python)
  - Large multidimensional array operations
- **SciPy** (Scientific Python)
  - Many efficient numerical routines such as routines for numerical integration and optimization
- **Pandas**
  - Data manipulation and data visualization
- **Matplotlib**
  - Data exploration and data visualization
- **Seaborn**
  - High-level data visualization library based on Matplotlib
- **Scikit-learn**
  - Machine learning and statistical modeling

Recall: Class and object

- Python is an object-oriented programming (OOP) language. Object-oriented programming is a programming paradigm based on the concept of objects.
- Class: A class is a blueprint to create objects.
- Object: An object is an instance of a class.

```
mystr = "hello world"
print(type(mystr))
```

Create a new object of class “str“.

```
<class 'str'>
```

```
mystr.upper()
```

Functions bound to objects are called methods.

```
'HELLO WORLD'
```

User-defined Class

- You can define your own classes.
  - Attributes: Variables of a class.
  - Methods: Functions of a class.

```
# Define a class
class Student:
    def __init__(self, school, name):
        self.school = school
        self.name = name

    def greeting(self):
        print("Hi, I'am a student at {}.".format(self.school))
```

attributes

method

```
# Create objects
student1 = Student("BI", "Anna")
student2 = Student("BI", "Lucas")
```

```
print(type(student1))
```

```
<class '__main__.Student'>
```

```
# Call the method
student1.greeting()
```

```
Hi, I'am a student at BI.
```

```
# Access the attribute
student1.name
```

```
'Anna'
```

- https://www.w3schools.com/python/python_classes.asp
- https://docs.python.org/3/tutorial/classes.html

Pandas

Pandas

- Pandas
  - An open-source package providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language.
- Installation
  - Installing pandas from a jupyter notebook

```
pip install pandas
```

  - Installing pandas from Anaconda Navigator
    - https://docs.anaconda.com/anaconda/navigator/tutorials/pandas/
- Import package
  - The "`as pd`" part means that we can write the shorthand "pd" instead of "pandas" when we use it later. In principle, you could use other names than "pd", but this is the common alias.

```
import pandas as pd
```


https://pandas.pydata.org/pandas-docs/stable/user_guide/cookbook.html#cookbook

Data structures

- **Python list**

```
mylist = [1,2,3,4,5]
print(mylist)
```

```
[1, 2, 3, 4, 5]
```

- **NumPy array**:
  - Numpy array can directly handle a mathematical operations.

```
import numpy as np
myarray = np.array([1,2,3,4,5])
print(myarray)
```

```
[1 2 3 4 5]
```

```
np.square(myarray)
```

```
array([ 1,  4,  9, 16, 25])
```

```
np.log(myarray)
```

```
array([0.        , 0.69314718, 1.09861229, 1.38629436, 1.60943791])
```

Pandas data structures

- **Pandas data structures (data type)**
  - **Series class:** A one-dimensional array-like structure.
    - A series contains an array of data and an associated array of index.

0	216
1	143
2	98
3	455
4	108

  - **DataFrame class:** A tabular, spreadsheet-like structure.
    - A DataFrame contains an ordered collection of columns, each of which can be a different data type (numeric, string, boolean, etc.).
    - A DataFrame has both a row and column index.

	High	Width	weight	group
0	20	20	0.1	A
1	45	30	0.8	C
2	54	43	1.5	C
3	25	15	2.3	B
4	18	34	0.2	A

DataFrame

	High
0	20
1	45
2	54
3	25
4	18

Series

	Width
0	20
1	30
2	43
3	15
4	34

Series

	weight
0	0.1
1	0.8
2	1.5
3	2.3
4	0.2

Series

	group
0	A
1	C
2	C
3	B
4	A

Series

Series - creation

- Create a series from a list

```
Series1 = pd.Series([4, 7, -5, 3])
Series1
```

```
0    4
1    7
2   -5
3    3
dtype: int64
```

index

- Get values and index

```
Series1.values
```

```
array([ 4,  7, -5,  3], dtype=int64)
```

The values are simply a NumPy array.

```
Series1.index
```

```
RangeIndex(start=0, stop=4, step=1)
```

RangeIndex is the default index type used by Series and DataFrame when no explicit index is provided by the user.

https://pandas.pydata.org/docs/reference/api/pandas.Series.html

Series - index

- Assign index

```
Series2 = pd.Series([4, 7, -5, 3], index = ["a","b","c","d"])
Series2
```

```
a    4
b    7
c   -5
d    3
dtype: int64
```

- Get index

```
Series2.index
```

```
Index(['a', 'b', 'c', 'd'], dtype='object')
```


https://pandas.pydata.org/docs/reference/api/pandas.Index.html

Series – subset selection

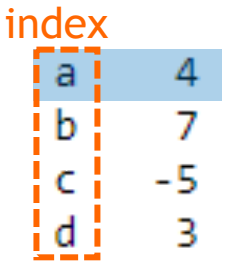
```
Series2 = pd.Series([4, 7, -5, 3], index = ["a","b","c","d"])
Series2
```

```
a    4
b    7
c   -5
d    3
dtype: int64
```

- Select a single value

```
Series2['a']
```

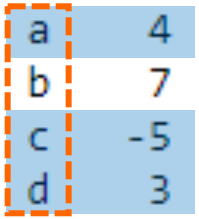
```
4
```



- Select a set of values

```
Series2[['c', 'a', 'd']]
```

```
c   -5
a    4
d    3
dtype: int64
```



Series - filter

- Use Boolean array to filter data.

```
Series2 > 2
```

```
a     True
b     True
c    False
d     True
dtype: bool
```

Boolean array

		>2
a	4	True
b	7	True
c	-5	False
d	3	True

```
Series2[Series2 > 2]
```

```
a    4
b    7
d    3
dtype: int64
```

Recall – logical operators

- Logical operators are used to combine boolean constraints.

Logical operator	Meaning
and	and
or	or
not	not

```
x = 2
y = 5
```

```
x == 2 and y == 5
```

```
True
```

```
x < 0 and y == 5
```

```
False
```

```
x < 0 or y == 5
```

```
True
```

Series - filter

- Pandas uses bitwise operators to combine conditions.

Bitwise operator	Meaning
&	and
|	or
~	not

```
(Series2 > 2) & (Series2 < 5)
```

```
a     True
b    False
c    False
d     True
dtype: bool
```

		>2		<5		
a	4	True		True		True
b	7	True	&	False	➡	False
c	-5	False		False		False
d	3	True		True		True

```
Series2[(Series2 > 2) & (Series2 < 5)]
```

```
a    4
d    3
dtype: int64
```

a	4	True
b	7	False
c	-5	False
d	3	True

https://www.geeksforgeeks.org/python-bitwise-operators/

Exercise

(A.1) Create a series with `score_list` as the value and `ID_list` as the index.

```
ID_list = ['S01', 'S02', 'S03', 'S04', 'S05']
score_list = [7.0, 5.5, 9.0, 5.0, 7.5]
```

(A.2) Select the data of students `'S01'` and `'S03'`.

(A.3) Select students with a score less than 6.

DataFrame - creation

- DataFrame: A tabular, spreadsheet-like structure.
- Create a DataFrame from a dictionary of equal-length lists.

```
data = {"state": ["Ohio","Ohio","Ohio","Nevada","Nevada","Nevada"],
        "year":[2000,2001,2002,2001,2002,2003],
        "pop":[1.5,1.7,3.6,2.4,2.9,3.2]}
```

```
df = pd.DataFrame(data, index = ["a","b","c","d","e","f"])
df
```

	state	year	pop
a	Ohio	2000	1.5
b	Ohio	2001	1.7
c	Ohio	2002	3.6
d	Nevada	2001	2.4
e	Nevada	2002	2.9
f	Nevada	2003	3.2

https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.html

DataFrame – labels and integer positions

- Axis -labels

```
df.index
```

```
Index(['a', 'b', 'c', 'd', 'e', 'f'], dtype='object')
```

```
df.columns
```

```
Index(['state', 'year', 'pop'], dtype='object')
```

	state	year	pop
a	Ohio	2000	1.5
b	Ohio	2001	1.7
c	Ohio	2002	3.6
d	Nevada	2001	2.4
e	Nevada	2002	2.9
f	Nevada	2003	3.2

Columns (label)

Index (label)

- Integer positions

	0	1	2
	state	year	pop
0	a	Ohio	2000	1.5
1	b	Ohio	2001	1.7
2	c	Ohio	2002	3.6
3	d	Nevada	[3,1]	2.4
4	e	Nevada	2002	2.9
5	f	Nevada	2003	3.2

Integer position

DataFrame – subset selection (1) by columns and rows

	state	year	pop
a	Ohio	2000	1.5
b	Ohio	2001	1.7
c	Ohio	2002	3.6
d	Nevada	2001	2.4
e	Nevada	2002	2.9
f	Nevada	2003	3.2

- Using axis-lable (**loc**).

```
df.loc[["b","c","d"], ["state","year"]]
```

	state	year
b	Ohio	2001
c	Ohio	2002
d	Nevada	2001

- Using integer position (**iloc**).

```
df.iloc[1:4, 0:2]
```

	state	year
b	Ohio	2001
c	Ohio	2002
d	Nevada	2001

DataFrame – subset selection (2) by columns or rows

	state	year	pop
a	Ohio	2000	1.5
b	Ohio	2001	1.7
c	Ohio	2002	3.6
d	Nevada	2001	2.4
e	Nevada	2002	2.9
f	Nevada	2003	3.2

- Using column names to select columns

```
df[["state","year"]]   # same as df.loc[:,["state","year"]]
```

	state	year
a	Ohio	2000
b	Ohio	2001
c	Ohio	2002
d	Nevada	2001
e	Nevada	2002
f	Nevada	2003

		state	year	pop
0	a	Ohio	2000	1.5
1	b	Ohio	2001	1.7
2	c	Ohio	2002	3.6
3	d	Nevada	2001	2.4
4	e	Nevada	2002	2.9
5	f	Nevada	2003	3.2

- Using integer positions to select rows

```
df[1:4]                # same as df.iloc[1:4,:]
```

	state	year	pop
b	Ohio	2001	1.7
c	Ohio	2002	3.6
d	Nevada	2001	2.4

DataFrame - filter

- Use Boolean array to filter data.

```
df[df.year > 2001]
```

	state	year	pop
2	Ohio	2002	3.6
4	Nevada	2002	2.9
5	Nevada	2003	3.2

```
df[df.state == "Ohio"]
```

	state	year	pop
0	Ohio	2000	1.5
1	Ohio	2001	1.7
2	Ohio	2002	3.6

	state	year	pop	df.year>2001
0	Ohio	2000	1.5	False
1	Ohio	2001	1.7	False
2	Ohio	2002	3.6	True
3	Nevada	2001	2.4	False
4	Nevada	2002	2.9	True
5	Nevada	2003	3.2	True

	state	year	pop	df.state=Ohio
0	Ohio	2000	1.5	True
1	Ohio	2001	1.7	True
2	Ohio	2002	3.6	True
3	Nevada	2001	2.4	False
4	Nevada	2002	2.9	False
5	Nevada	2003	3.2	False

DataFrame - filter

- Use bitwise operators to combine conditions

```
df[(df.state == "Ohio") & (df.year > 2000)]
```

	state	year	pop
b	Ohio	2001	1.7
c	Ohio	2002	3.6

	state	year	pop
0	Ohio	2000	1.5
1	Ohio	2001	1.7
2	Ohio	2002	3.6
3	Nevada	2001	2.4
4	Nevada	2002	2.9
5	Nevada	2003	3.2

Exercise

(B.1) Create a dictionary with `company_name, profit, assets` as the keys and `list1, list2, list3` as the values. Print out the dictionary.

```
list1 = ['JPMorgan Chase','Apple','Bank of America','Amazon','Microsoft']
list2 = [40.4, 63.9, 17.9, 21.3, 51.3]
list3 = [3689.3, 354.1, 2832.2, 321.2, 304.1]
```

(B.2) Create a dataframe named `mydf` based on the dictionary defined in (B.1) and use `a, b, c, d, e` as index. Print out the dataframe.

(B.3) Use `loc` to select a subset as follows.

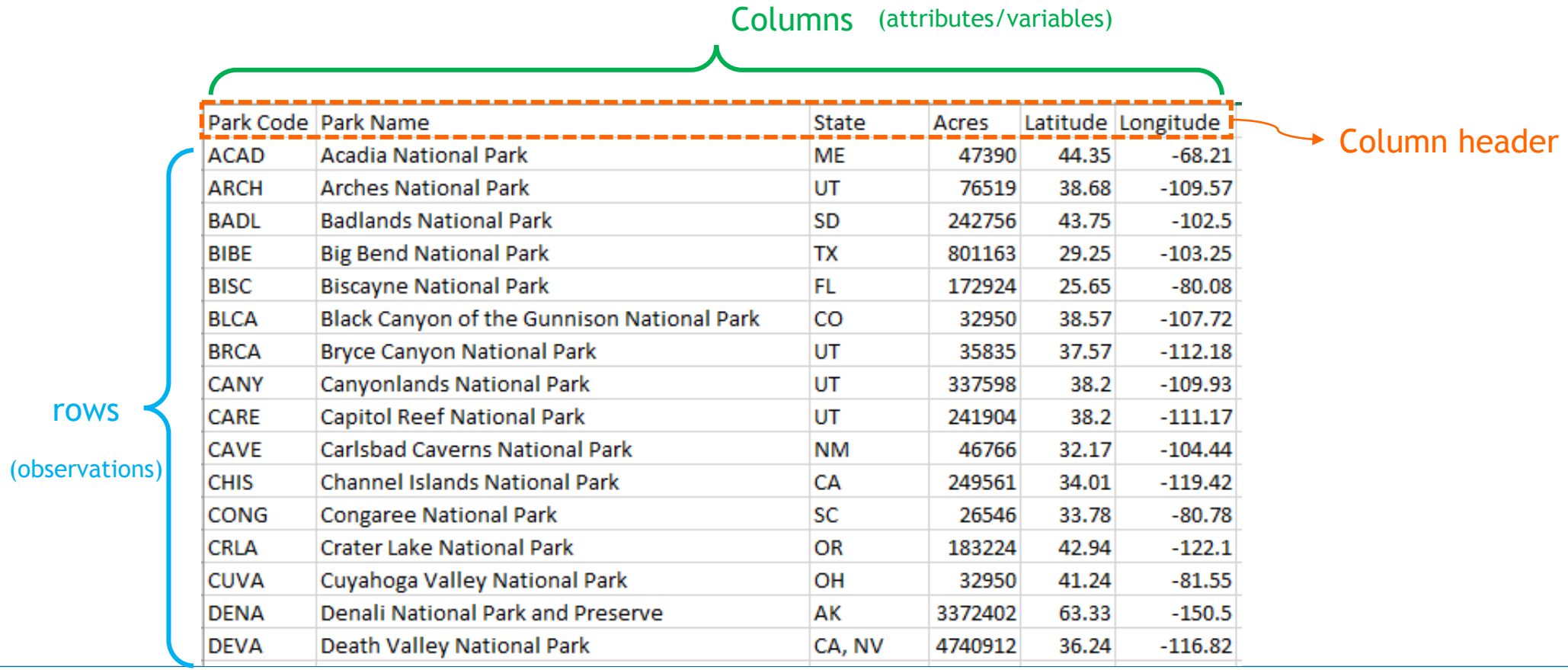
company name	assets
Apple	354.1
Amazon	321.2
Microsoft	304.1

(B.4) Use `iloc` to select the same subset.

Read data from file to DataFrame

Dataset

- A dataset is a collection of data with a defined structure.



Park Code	Park Name	State	Acres	Latitude	Longitude
ACAD	Acadia National Park	ME	47390	44.35	-68.21
ARCH	Arches National Park	UT	76519	38.68	-109.57
BADL	Badlands National Park	SD	242756	43.75	-102.5
BIBE	Big Bend National Park	TX	801163	29.25	-103.25
BISC	Biscayne National Park	FL	172924	25.65	-80.08
BLCA	Black Canyon of the Gunnison National Park	CO	32950	38.57	-107.72
BRCA	Bryce Canyon National Park	UT	35835	37.57	-112.18
CANY	Canyonlands National Park	UT	337598	38.2	-109.93
CARE	Capitol Reef National Park	UT	241904	38.2	-111.17
CAVE	Carlsbad Caverns National Park	NM	46766	32.17	-104.44
CHIS	Channel Islands National Park	CA	249561	34.01	-119.42
CONG	Congaree National Park	SC	26546	33.78	-80.78
CRLA	Crater Lake National Park	OR	183224	42.94	-122.1
CUVA	Cuyahoga Valley National Park	OH	32950	41.24	-81.55
DENA	Denali National Park and Preserve	AK	3372402	63.33	-150.5
DEVA	Death Valley National Park	CA, NV	4740912	36.24	-116.82

Download csv file

- Download the dataset from itslearning
  - Resources/Part2: Data Extraction and Visualization/Dataset/**parks.csv** (US national park data)

Open in Excel

	A	B	C	D	E	F
1	Park Code	Park Name	State	Acres	Latitude	Longitude
2	ACAD	Acadia National Pa	ME	47390	44.35	-68.21
3	ARCH	Arches National Pa	UT	76519	38.68	-109.57
4	BADL	Badlands National	SD	242756	43.75	-102.5
5	BIBE	Big Bend National F	TX	801163	29.25	-103.25
6	BISC	Biscayne National F	FL	172924	25.65	-80.08
7	BLCA	Black Canyon of the	CO	32950	38.57	-107.72
8	BRCA	Bryce Canyon Natic	UT	35835	37.57	-112.18
9	CANY	Canyonlands Natio	UT	337598	38.2	-109.93
10	CARE	Capitol Reef Natior	UT	241904	38.2	-111.17
11	CAVE	Carlsbad Caverns N	NM	46766	32.17	-104.44
12	CHIS	Channel Islands Na	CA	249561	34.01	-119.42
13	CONG	Congaree National	SC	26546	33.78	-80.78
14	CRLA	Crater Lake Nationa	OR	183224	42.94	-122.1
15	CUVA	Cuyahoga Valley Na	OH	32950	41.24	-81.55
16	DENA	Denali National Par	AK	3372402	63.33	-150.5
17	DEVA	Death Valley Natio	CA, NV	4740912	36.24	-116.82
18	DRTO	Dry Tortugas Natior	FL	64701	24.63	-82.87
19	EVER	Everglades Nationa	FL	1508538	25.32	-80.93
20	GAAR	Gates Of The Arctic	AK	7523898	67.78	-153.3
21	GLAC	Glacier National Pa	MT	1013572	48.8	-114
22	GLBA	Glacier Bay Nationa	AK	3224840	58.5	-137
23	GRBA	Great Basin Nationa	NV	77180	38.98	-114.3
24	GRCA	Grand Canyon Nati	AZ	1217403	36.06	-112.14

Open in NotePad

parks.csv

```
Park Code,Park Name,State,Acres,Latitude,Longitude
ACAD,Acadia National Park,ME,47390,44.35,-68.21
ARCH,Arches National Park,UT,76519,38.68,-109.57
BADL,Badlands National Park,SD,242756,43.75,-102.5
BIBE,Big Bend National Park,TX,801163,29.25,-103.25
BISC,Biscayne National Park,FL,172924,25.65,-80.08
BLCA,Black Canyon of the Gunnison National Park,CO,32950,38.57,-107.72
BRCA,Bryce Canyon National Park,UT,35835,37.57,-112.18
CANY,Canyonlands National Park,UT,337598,38.2,-109.93
CARE,Capitol Reef National Park,UT,241904,38.2,-111.17
CAVE,Carlsbad Caverns National Park,NM,46766,32.17,-104.44
CHIS,Channel Islands National Park,CA,249561,34.01,-119.42
CONG,Congaree National Park,SC,26546,33.78,-80.78
CRLA,Crater Lake National Park,OR,183224,42.94,-122.1
CUVA,Cuyahoga Valley National Park,OH,32950,41.24,-81.55
DENA,Denali National Park and Preserve,AK,3372402,63.33,-150.5
DEVA,Death Valley National Park,"CA, NV",4740912,36.24,-116.82
DRTO,Dry Tortugas National Park,FL,64701,24.63,-82.87
EVER,Everglades National Park,FL,1508538,25.32,-80.93
GAAR,Gates Of The Arctic National Park and Preserve,AK,7523898,67.78,-153
GLAC,Glacier National Park,MT,1013572,48.8,-114
GLBA,Glacier Bay National Park and Preserve,AK,3224840,58.5,-137
GRBA,Great Basin National Park,NV,77180,38.98,-114.3
GRCA,Grand Canyon National Park,AZ,1217403,36.06,-112.14
GRSA,Great Sand Dunes National Park and Preserve,CO,42984,37.73,-105.51
GRSM,Great Smoky Mountains National Park,"TN, NC",521490,35.68,-83.53
GRTE,Grand Teton National Park,WY,309995,43.73,-110.8
GUMO,Guadalupe Mountains National Park,TX,86416,31.92,-104.87
HALE,Haleakala National Park,HI,29094,20.72,-156.17
HAVO,Hawaii Volcanoes National Park,HI,323431,19.38,-155.2
HOSP,Hot Springs National Park,AR,5550,34.51,-93.05
ISRO,Isle Royale National Park,MI,571790,48.1,-88.55
```

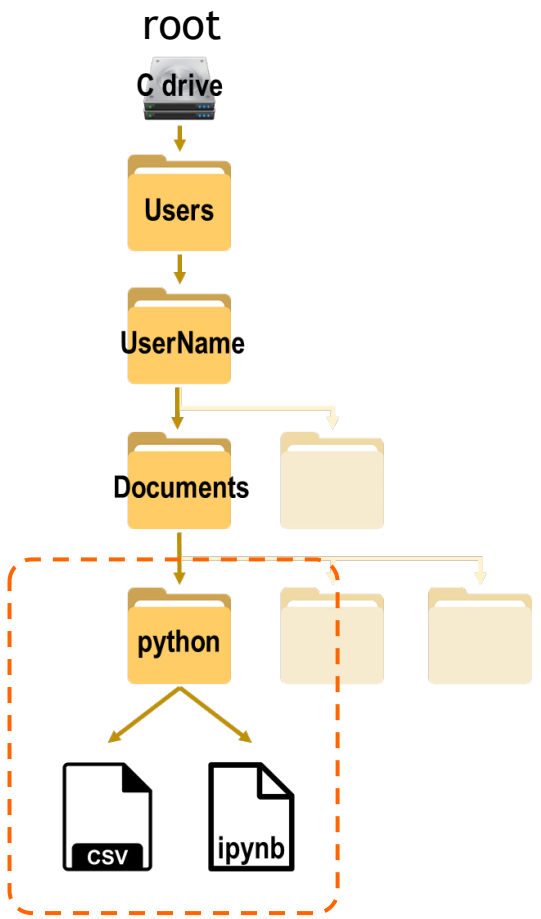
- A .csv file ("comma separated value") is a file where the data is stored in a text format, separated by commas (or other separation marks such as space or semi-colon).

Absolute path and relative path

- An absolute path is defined as specifying the location of a file from the root directory. An absolute path is a complete path.

windows
```
df = pd.read_csv("C:/Users/UserName/Documents/python/park.csv")
```

Mac
```
df = pd.read_csv("/Users/UserName/Documents/python/park.csv")
```



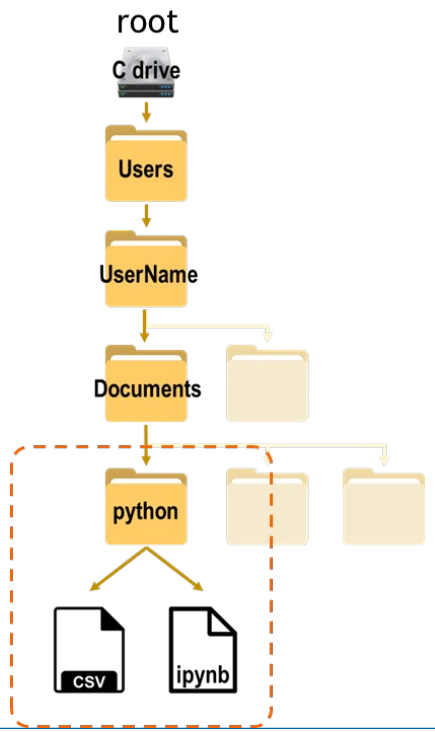
- A relative path is defined as the location related to the current working file. Pandas can find the file from where your notebook is running.

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

Read csv file into a jupyter notebook

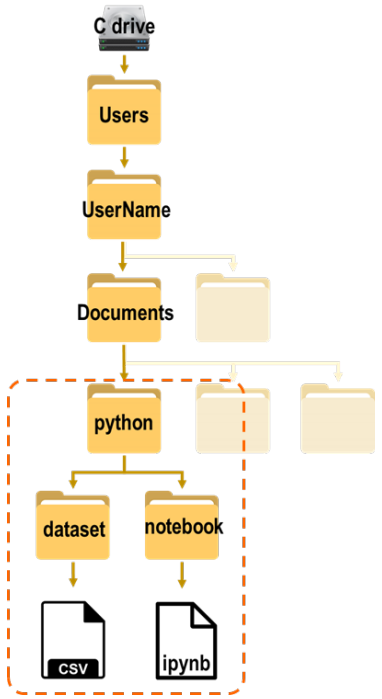
Case1: If you put the csv file in the same folder as jupyter notebook.

```
park_df = pd.read_csv('parks.csv')
```



Case2: If you put csv file in a different folder.

```
park_df = pd.read_csv('../dataset/parks.csv')
```



- You will have to specify the location of the file on your computer. The location is relative to where you will find this notebook itself.
- In this example, the file "parks.csv" is stored one folder back (../), then inside the folder “dataset".

View data

- View the first N rows

```
park_df.head(5)
```

	Park Code	Park Name	State	Acres	Latitude	Longitude
0	ACAD	Acadia National Park	ME	47390	44.35	-68.21
1	ARCH	Arches National Park	UT	76519	38.68	-109.57
2	BADL	Badlands National Park	SD	242756	43.75	-102.50
3	BIBE	Big Bend National Park	TX	801163	29.25	-103.25
4	BISC	Biscayne National Park	FL	172924	25.65	-80.08

- Get the number of rows and columns

```
park_df.shape
```

```
(56, 6)
```

number of rows

number of columns

Subset selection – loc and iloc

- Using axis-lable (`loc`).

```
park_df.loc[[10,11,12,13,14],["Park Code","Park Name"]]
```

	Park Code	Park Name
10	CHIS	Channel Islands National Park
11	CONG	Congaree National Park
12	CRLA	Crater Lake National Park
13	CUVA	Cuyahoga Valley National Park
14	DENA	Denali National Park and Preserve

- Using integer position (`iloc`).

```
park_df.iloc[10:15,0:2]
```

	Park Code	Park Name
10	CHIS	Channel Islands National Park
11	CONG	Congaree National Park
12	CRLA	Crater Lake National Park
13	CUVA	Cuyahoga Valley National Park
14	DENA	Denali National Park and Preserve

Filter

- One condition

```
park_df[park_df.State == "UT"]
```

	Park Code	Park Name	State	Acres	Latitude	Longitude
1	ARCH	Arches National Park	UT	76519	38.68	-109.57
6	BRCA	Bryce Canyon National Park	UT	35835	37.57	-112.18
7	CANY	Canyonlands National Park	UT	337598	38.20	-109.93
8	CARE	Capitol Reef National Park	UT	241904	38.20	-111.17
55	ZION	Zion National Park	UT	146598	37.30	-113.05

- Multiple conditions

```
park_df[(park_df.State == "UT") & (park_df.Acres > 50000)]
```

	Park Code	Park Name	State	Acres	Latitude	Longitude
1	ARCH	Arches National Park	UT	76519	38.68	-109.57
7	CANY	Canyonlands National Park	UT	337598	38.20	-109.93
8	CARE	Capitol Reef National Park	UT	241904	38.20	-111.17
55	ZION	Zion National Park	UT	146598	37.30	-113.05

Filter

- If we want to evaluate many "or" expressions, we can use `isin`.

```
park_df[park_df.State.isin(['WA', 'OR', 'CA'])]
```

	Park Code	Park Name	State	Acres	Latitude	Longitude
10	CHIS	Channel Islands National Park	CA	249561	34.01	-119.42
12	CRLA	Crater Lake National Park	OR	183224	42.94	-122.10
31	JOTR	Joshua Tree National Park	CA	789745	33.79	-115.90
36	LAVO	Lassen Volcanic National Park	CA	106372	40.49	-121.51
39	MORA	Mount Rainier National Park	WA	235625	46.85	-121.75
40	NOCA	North Cascades National Park	WA	504781	48.70	-121.20
41	OLYM	Olympic National Park	WA	922651	47.97	-123.50
43	PINN	Pinnacles National Park	CA	26606	36.48	-121.16
44	REDW	Redwood National Park	CA	112512	41.30	-124.00
47	SEKI	Sequoia and Kings Canyon National Parks	CA	865952	36.43	-118.68
54	YOSE	Yosemite National Park	CA	761266	37.83	-119.50

Same as

```
park_df[(park_df.State =='WA')|(park_df.State =='OR')|(park_df.State == 'CA')]
```

Create a subset

- Assign the returned dataframe to a new variable.

```
park_west_df = park_df[park_df.State.isin(['WA', 'OR', 'CA'])]
park_west_df
```

	Park Code	Park Name	State	Acres	Latitude	Longitude
10	CHIS	Channel Islands National Park	CA	249561	34.01	-119.42
12	CRLA	Crater Lake National Park	OR	183224	42.94	-122.10
31	JOTR	Joshua Tree National Park	CA	789745	33.79	-115.90
36	LAVO	Lassen Volcanic National Park	CA	106372	40.49	-121.51
39	MORA	Mount Rainier National Park	WA	235625	46.85	-121.75
40	NOCA	North Cascades National Park	WA	504781	48.70	-121.20
41	OLYM	Olympic National Park	WA	922651	47.97	-123.50
43	PINN	Pinnacles National Park	CA	26606	36.48	-121.16
44	REDW	Redwood National Park	CA	112512	41.30	-124.00
47	SEKI	Sequoia and Kings Canyon National Parks	CA	865952	36.43	-118.68
54	YOSE	Yosemite National Park	CA	761266	37.83	-119.50

Exercise

(C.1) Read the csv file `diabetes.csv` using pandas. Display the first 10 rows.

(C.2) What is the number of rows and columns in this data set?

(C.3) Select (display) column `BloodPressure` and column `BMI`.

(C.4) Select rows with `BMI` greater than 50.

(C.5) Select rows with either `BMI` greater than 50 or `BloodPressure` greater than 110.

Descriptive statistics - info

- A summary of a DataFrame
  - Index, data type of each columns, number of non-null values, and memory usage.

```
park_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 56 entries, 0 to 55
Data columns (total 6 columns):
 #   Column      Non-Null Count  Dtype  
---  ------      --------------  -----  
 0   Park Code   56 non-null     object 
 1   Park Name   56 non-null     object 
 2   State       56 non-null     object 
 3   Acres       56 non-null     int64  
 4   Latitude    56 non-null     float64
 5   Longitude   56 non-null     float64
dtypes: float64(2), int64(1), object(3)
memory usage: 2.8+ KB
```

In Pandas, string data is always stored with an object dtype.

Descriptive statistics - pandas data types

- Pandas `dtype` mapping

Pandas dtype	Python type	Usage
int64	int	Integer numbers
float64	float	Floating point numbers
object	str or mixed	Text or mixed numeric and non-numeric values
bool	bool	True/False values
datetime64	datetime	Date and time values
timedelta[ns]	--	Differences between two datetimes
category	--	Finite list of text values

numerical data (quantitative)

categorical data (qualitative data)

- Numpy data type (dtype) https://numpy.org/doc/stable/reference/arrays.dtypes.html
- Example: int64 (64-bit integers): -9223372036854775808 to 9223372036854775807

Descriptive statistics – categorical data and numerical data

- **Categorical data** describes characteristics or groups.
  - Gender (Male, Female)
  - Product types (Wood, Plastic, Metal)
  - Risk level (Low, Medium, High)
  - Days of the week (Monday, Tuesday, Wednesday)
  - Months of the year (January, February, March)

- **Numerical data** expresses information in the form of numbers.
  - Number of customers (25, 19, 35,...)
  - Prices of products (200, 150, 80,...)
  - Interest rate (0.5, 1.4, 2.3,...)

Descriptive statistics – change data type

- Use `astype()` to change the data type (dtype).

```
park_df.Acres = park_df.Acres.astype(float)
park_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 56 entries, 0 to 55
Data columns (total 6 columns):
 #   Column      Non-Null Count  Dtype  
---  ------      --------------  -----  
 0   Park Code   56 non-null     object 
 1   Park Name   56 non-null     object 
 2   State       56 non-null     object 
 3   Acres       56 non-null     float64
 4   Latitude    56 non-null     float64
 5   Longitude   56 non-null     float64
dtypes: float64(3), object(3)
memory usage: 2.8+ KB
```

Descriptive statistics - describe

- Descriptive statistics of numerical columns.

```
park_df.describe()
```

	Acres	Latitude	Longitude
count	5.600000e+01	56.000000	56.000000
mean	9.279291e+05	41.233929	-113.234821
std	1.709258e+06	10.908831	22.440287
min	5.550000e+03	19.380000	-159.280000
25%	6.901050e+04	35.527500	-121.570000
50%	2.387645e+05	38.550000	-110.985000
75%	8.173602e+05	46.880000	-103.400000
max	8.323148e+06	67.780000	-68.210000

- https://en.wikipedia.org/wiki/Standard_deviation
- https://en.wikipedia.org/wiki/Quartile
- https://en.wikipedia.org/wiki/Scientific_notation

Descriptive statistics – value_counts

- Use `value_counts()` to count unique values in a column.

```
park_df.State.value_counts()
```

```
AK            8
CA            7
UT            5
CO            4
FL            3
WA            3
AZ            3
TX            2
SD            2
HI            2
WY            1
MN            1
ME            1
ND            1
OR            1
SC            1
CA, NV        1
MT            1
NM            1
NV            1
MI            1
VA            1
KY            1
WY, MT, ID    1
OH            1
TN, NC        1
AR            1
Name: State, dtype: int64
```

Exercise

Exercise.D

(D.1) Use the same dataset `diabetes.csv` in (C.1). What is the data type of each variable?

(D.2) Change the data type of `Outcome` to `object`.

(D.3) Get the average of variable `Age`.

(D.4) The variable `Outcome` records whether the person has diabetes: 0 means non-diabetic, 1 means diabetes. How many people in this dataset have diabetes?

Appendix

Packages – pip install

- Anaconda Navigator ➢ Environments ➢ base(root) ➢ Open Terminal