Data management using Pandas

Data management is a crucial component to statistical analysis and data science work.

This notebook will show you how to import, view, undertand, and manage your data using the Pandas data processing library, i.e., the notebook will demonstrates how to read a dataset into Python, and obtain a basic understanding of its content.

Note that **Python** by itself is a general-purpose programming language and does not provide high-level data processing capabilities. The **Pandas** library was developed to meet this need. **Pandas** is the most popular Python library for data manipulation, and we will use it extensively in this course. **Pandas** provides high-performance, easy-to-use data structures and data analysis tools.

The main data structure that **Pandas** works with is called a **Data Frame**. This is a two-dimensional table of data in which the rows typically represent cases and the columns represent variables (e.g. data used in this tutorial). Pandas also has a one-dimensional data structure called a **Series** that we will encounter when accesing a single column of a Data Frame.

Pandas has a variety of functions named `read_xxx` for reading data in different formats. Right now we will focus on reading `csv` files, which stands for comma-separated values. However the other file formats include `excel`, `json`, and `sql`.

There are many other options to `read_csv` that are very useful. For example, you would use the option `sep='\t'` instead of the default `sep=','` if the fields of your data file are delimited by tabs instead of commas. See here for the full documentation for `read_csv`.

Acknowledgments

- The dataset used in this tutorial is from https://www.coursera.org/ from the course "Understanding and Visualizing Data with Python" by University of Michigan

Importing libraries

In [106…
```
# Import the packages that we will be using
import pandas as pd
```

Importing data

In [107…
```
# Define where you are running the code: colab or local
```

In [108…
```
# url string that hosts our .csv file

# Read the .csv file and store it as a pandas Data Frame
df = pd.read_csv(r'C:\VisualProyectsPC\TC1002S\llddl\23\TC1002S\NotebooksStudents\A01642529\Iris
```

If we want to print the information about th output object type we would simply type the following: type(df)

In [109…
```
type(df)
```

Out[109…
```
pandas.core.frame.DataFrame
```

Exploring the content of the data set

Use the `shape` method to determine the numbers of rows and columns in a data frame. This can be used to confirm that we have actually obtained the data the we are expecting.

Based on what we see below, the data set being read here has N_r rows, corresponding to N_r observations, and N_c columns, corresponding to N_c variables in this particular data file.

In [110…
```
df.shape
```

Out[110…
```
(150, 6)
```

In [111…
```
# Get number of rows
Nrows = df.shape[0]

print('The Number of rows is ' + str(Nrows))
```

```
The Number of rows is 150
```

In [112…
```
# Get number of cols
Ncols = df.shape[1]

print('The Number of columns is ' + str(Ncols))
```

```
The Number of columns is 6
```

If we want to show the entire data frame we would simply write the following:

In [113…
```
df
```

Out[113…

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	1	5.1	3.5	1.4	0.2	Iris-setosa
1	2	4.9	3.0	1.4	0.2	Iris-setosa
2	3	4.7	3.2	1.3	0.2	Iris-setosa
3	4	4.6	3.1	1.5	0.2	Iris-setosa
4	5	5.0	3.6	1.4	0.2	Iris-setosa
...	...	...	...	...	...	...
145	146	6.7	3.0	5.2	2.3	Iris-virginica
146	147	6.3	2.5	5.0	1.9	Iris-virginica
147	148	6.5	3.0	5.2	2.0	Iris-virginica
148	149	6.2	3.4	5.4	2.3	Iris-virginica
149	150	5.9	3.0	5.1	1.8	Iris-virginica

150 rows × 6 columns

As you can see, we have a 2-Dimensional object where each row is an independent observation and each coloum is a variable.

Now, use the the `head()` function to show the first 5 rows of our data frame

In [114…

```
df.head(5)
```

Out[114…

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	1	5.1	3.5	1.4	0.2	Iris-setosa
1	2	4.9	3.0	1.4	0.2	Iris-setosa
2	3	4.7	3.2	1.3	0.2	Iris-setosa
3	4	4.6	3.1	1.5	0.2	Iris-setosa
4	5	5.0	3.6	1.4	0.2	Iris-setosa

Also, you can use the the `tail()` function to show the last 5 rows of our data frame

In [115…

```
df.tail(5)
```

Out[115…

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
145	146	6.7	3.0	5.2	2.3	Iris-virginica
146	147	6.3	2.5	5.0	1.9	Iris-virginica
147	148	6.5	3.0	5.2	2.0	Iris-virginica
148	149	6.2	3.4	5.4	2.3	Iris-virginica
149	150	5.9	3.0	5.1	1.8	Iris-virginica

The columns in a Pandas data frame have names, to see the names, use the `columns` method:

To gather more information regarding the data, we can view the column names with the following function:

In [116…
```
df.columns
```

Out[116…
```
Index(['Id', 'SepalLengthCm', 'SepalWidthCm', 'PetalLengthCm', 'PetalWidthCm',
       'Species'],
      dtype='object')
```

Be aware that every variable in a Pandas data frame has a data type. There are many different data types, but most commonly you will encounter floating point values (real numbers), integers, strings (text), and date/time values. When Pandas reads a text/csv file, it guesses the data types based on what it sees in the first few rows of the data file. Usually it selects an appropriate type, but occasionally it does not. To confirm that the data types are consistent with what the variables represent, inspect the `dtypes` attribute of the data frame.

In [117…
```
df.dtypes
```

Out[117…
```
Id                int64
SepalLengthCm    float64
SepalWidthCm     float64
PetalLengthCm    float64
PetalWidthCm     float64
Species           object
dtype: object
```

Summary statistics, which include things like the mean, min, and max of the data, can be useful to get a feel for how large some of the variables are and what variables may be the most important.

In [118…
```
# Summary statistics for the quantitative variables
df.describe()
```

Out[118…

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm
count	150.000000	150.000000	150.000000	150.000000	150.000000
mean	75.500000	5.843333	3.054000	3.758667	1.198667
std	43.445368	0.828066	0.433594	1.764420	0.763161
min	1.000000	4.300000	2.000000	1.000000	0.100000
25%	38.250000	5.100000	2.800000	1.600000	0.300000
50%	75.500000	5.800000	3.000000	4.350000	1.300000
75%	112.750000	6.400000	3.300000	5.100000	1.800000
max	150.000000	7.900000	4.400000	6.900000	2.500000

In [119…

```
# Drop observations with NaN values
df.dropna()
df.shape
```

Out[119…

```
(150, 6)
```

It is also possible to get statistics on the entire data frame or a column as follows

- `df.mean()` Returns the mean of all columns
- `df.corr()` Returns the correlation between columns in a data frame
- `df.count()` Returns the number of non-null values in each data frame column
- `df.max()` Returns the highest value in each column
- `df.min()` Returns the lowest value in each column
- `df.median()` Returns the median of each column
- `df.std()` Returns the standard deviation of each column

In [120…

```
df.SepalLengthCm.mean()
```

Out[120…

```
np.float64(5.843333333333334)
```

In [121…

```
df.max()
```

Out[121…

```
Id                            150
SepalLengthCm                 7.9
SepalWidthCm                  4.4
PetalLengthCm                 6.9
PetalWidthCm                  2.5
Species          Iris-virginica
dtype: object
```

In [122…

```
df.min()
```

Out[122…

```
Id                            1
SepalLengthCm               4.3
SepalWidthCm                2.0
PetalLengthCm               1.0
PetalWidthCm                0.1
Species           Iris-setosa
dtype: object
```

In [123…
```
df.median(numeric_only=True)
```

Out[123…
```
Id              75.50
SepalLengthCm    5.80
SepalWidthCm     3.00
PetalLengthCm    4.35
PetalWidthCm     1.30
dtype: float64
```

In [124…
```
df.corr(numeric_only=True)
```

Out[124…

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm
Id	1.000000	0.716676	-0.397729	0.882747	0.899759
SepalLengthCm	0.716676	1.000000	-0.109369	0.871754	0.817954
SepalWidthCm	-0.397729	-0.109369	1.000000	-0.420516	-0.356544
PetalLengthCm	0.882747	0.871754	-0.420516	1.000000	0.962757
PetalWidthCm	0.899759	0.817954	-0.356544	0.962757	1.000000

In [125…
```
df.std(numeric_only=True)
```

Out[125…
```
Id              43.445368
SepalLengthCm    0.828066
SepalWidthCm     0.433594
PetalLengthCm    1.764420
PetalWidthCm     0.763161
dtype: float64
```

In [126…
```
df.count()
```

Out[126…
```
Id              150
SepalLengthCm   150
SepalWidthCm    150
PetalLengthCm   150
PetalWidthCm    150
Species         150
dtype: int64
```

How to write a data frame to a File

To save a file with your data simply use the `to_csv` attribute

Examples:

- df.to_csv('myDataFrame.csv')
- df.to_csv('myDataFrame.csv', sep='\t')

In [127…
```
df.to_csv('IrisOut.csv')
```

Rename columns

To change the name of a colum use the `rename` attribute

Example:

df = df.rename(columns={"Age": "Edad"})

df.head()

In [128…
```
df = df.rename(columns={"SepalLengthCm":"SepalLCM","SepalWidthCm":"SepalWCM"})
df.head()
```

Out[128…

	Id	SepalLCM	SepalWCM	PetalLengthCm	PetalWidthCm	Species
0	1	5.1	3.5	1.4	0.2	Iris-setosa
1	2	4.9	3.0	1.4	0.2	Iris-setosa
2	3	4.7	3.2	1.3	0.2	Iris-setosa
3	4	4.6	3.1	1.5	0.2	Iris-setosa
4	5	5.0	3.6	1.4	0.2	Iris-setosa

In [129…
```
# Back to the original name
df = df.rename(columns={"SepalLCM":"SepalLengthCm","SepalWCM":"SepalWidthCm"})
df.head()
```

Out[129…

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	1	5.1	3.5	1.4	0.2	Iris-setosa
1	2	4.9	3.0	1.4	0.2	Iris-setosa
2	3	4.7	3.2	1.3	0.2	Iris-setosa
3	4	4.6	3.1	1.5	0.2	Iris-setosa
4	5	5.0	3.6	1.4	0.2	Iris-setosa

Selection of colums

As discussed above, a Pandas data frame is a rectangular data table, in which the rows represent observations or samples and the columns represent variables. One common manipulation of a data frame is to extract the data for one case or for one variable. There are several ways to do this, as shown below.

To extract all the values for one column (variable), use one of the following alternatives.

In [130…
```
df[["Id", "Species"]]
```

Out[130…

	Id	Species
0	1	Iris-setosa
1	2	Iris-setosa
2	3	Iris-setosa
3	4	Iris-setosa
4	5	Iris-setosa
...	...	...
145	146	Iris-virginica
146	147	Iris-virginica
147	148	Iris-virginica
148	149	Iris-virginica
149	150	Iris-virginica

150 rows × 2 columns

Slicing a data set

As discussed above, a Pandas data frame is a rectangular data table, in which the rows represent cases and the columns represent variables. One common manipulation of a data frame is to extract the data for one observation or for one variable. There are several ways to do this, as shown below.

Lets say we would like to splice our data frame and select only specific portions of our data. There are three different ways of doing so.

1. .loc()
2. .iloc()
3. .ix()

We will cover the .loc() and .iloc() splicing functions.

The attibute **.loc()** uses labels/column names, in specific, it takes two single/list/range operator separated by ',', the first one indicates the rows and the second one indicates columns.

In [131…

```
# Return all observations of CWDistance
# Select few rows for multiple columns, ["CWDistance", "Height", "Wingspan"]
df.loc[4:9, ["Id", "SepalWidthCm", "Species"]]
```

Out[131…

	Id	SepalWidthCm	Species
4	5	3.6	Iris-setosa
5	6	3.9	Iris-setosa
6	7	3.4	Iris-setosa
7	8	3.4	Iris-setosa
8	9	2.9	Iris-setosa
9	10	3.1	Iris-setosa

The attribute **iloc()** is an integer based slicing.

In [132…
```
df.iloc[:, 3:5]
```

Out[132…

	PetalLengthCm	PetalWidthCm
0	1.4	0.2
1	1.4	0.2
2	1.3	0.2
3	1.5	0.2
4	1.4	0.2
…	…	…
145	5.2	2.3
146	5.0	1.9
147	5.2	2.0
148	5.4	2.3
149	5.1	1.8

150 rows × 2 columns

Get unique existing values

List unique values in the one of the columns

df.Gender.unique()

In [133…
```
# List unique values in the df['Gender'] column
df.Species.unique()
```

Out[133…
```
array(['Iris-setosa', 'Iris-versicolor', 'Iris-virginica'], dtype=object)
```

In [134…
```
# Lets explore df["GenderGroup] as well
df.PetalLengthCm.unique()
```

Out[134…
```
array([1.4, 1.3, 1.5, 1.7, 1.6, 1.1, 1.2, 1. , 1.9, 4.7, 4.5, 4.9, 4. ,
       4.6, 3.3, 3.9, 3.5, 4.2, 3.6, 4.4, 4.1, 4.8, 4.3, 5. , 3.8, 3.7,
       5.1, 3. , 6. , 5.9, 5.6, 5.8, 6.6, 6.3, 6.1, 5.3, 5.5, 6.7, 6.9,
       5.7, 6.4, 5.4, 5.2])
```

Filter, Sort and Groupby

With **Filter** you can use different conditions to filter columns. For example, df[df[year] > 1984] would give you only the column year is greater than 1984. You can use & (and) or | (or) to add different conditions to your filtering. This is also called boolean filtering.

df[df["Height"] >= 70]

In [135…
```
df[df["PetalLengthCm"] >= 3]
```

Out[135…

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
50	51	7.0	3.2	4.7	1.4	Iris-versicolor
51	52	6.4	3.2	4.5	1.5	Iris-versicolor
52	53	6.9	3.1	4.9	1.5	Iris-versicolor
53	54	5.5	2.3	4.0	1.3	Iris-versicolor
54	55	6.5	2.8	4.6	1.5	Iris-versicolor
…	…	…	…	…	…	…
145	146	6.7	3.0	5.2	2.3	Iris-virginica
146	147	6.3	2.5	5.0	1.9	Iris-virginica
147	148	6.5	3.0	5.2	2.0	Iris-virginica
148	149	6.2	3.4	5.4	2.3	Iris-virginica
149	150	5.9	3.0	5.1	1.8	Iris-virginica

100 rows × 6 columns

With **Sort** is possible to sort values in a certain column in an ascending order using `df.sort_values("ColumnName")` or in descending order using `df.sort_values(ColumnName, ascending=False)`.

Furthermore, it's possible to sort values by Column1Name in ascending order then Column2Name in descending order by using `df.sort_values([Column1Name,Column2Name],ascending=[True,False])`

df.sort_values("Height")

df.sort_values("Height",ascending=False)

In [136…
```
df.sort_values("PetalLengthCm",ascending=True)
```

Out[136…

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
22	23	4.6	3.6	1.0	0.2	Iris-setosa
13	14	4.3	3.0	1.1	0.1	Iris-setosa
14	15	5.8	4.0	1.2	0.2	Iris-setosa
35	36	5.0	3.2	1.2	0.2	Iris-setosa
16	17	5.4	3.9	1.3	0.4	Iris-setosa
…	…	…	…	…	…	…
131	132	7.9	3.8	6.4	2.0	Iris-virginica
105	106	7.6	3.0	6.6	2.1	Iris-virginica
117	118	7.7	3.8	6.7	2.2	Iris-virginica
122	123	7.7	2.8	6.7	2.0	Iris-virginica
118	119	7.7	2.6	6.9	2.3	Iris-virginica

150 rows × 6 columns

The attribute **Groupby** involves splitting the data into groups based on some criteria, applying a function to each group independently and combining the results into a data structure. df.groupby(col) returns a groupby object for values from one column while df.groupby([col1,col2]) returns a groupby object for values from multiple columns.

df.groupby(['Gender'])

In [137…
```
df.groupby(['Species']).size()
```

Out[137…
```
Species
Iris-setosa        50
Iris-versicolor    50
Iris-virginica     50
dtype: int64
```

Size of each group

In [138…
```
df.groupby(['SepalWidthCm','Species']).size()
```

Out[138…

```
SepalWidthCm  Species
2.0           Iris-versicolor     1
2.2           Iris-versicolor     2
              Iris-virginica      1
2.3           Iris-setosa         1
              Iris-versicolor     3
2.4           Iris-versicolor     3
2.5           Iris-versicolor     4
              Iris-virginica      4
2.6           Iris-versicolor     3
              Iris-virginica      2
2.7           Iris-versicolor     5
              Iris-virginica      4
2.8           Iris-versicolor     6
              Iris-virginica      8
2.9           Iris-setosa         1
              Iris-versicolor     7
              Iris-virginica      2
3.0           Iris-setosa         6
              Iris-versicolor     8
              Iris-virginica     12
3.1           Iris-setosa         5
              Iris-versicolor     3
              Iris-virginica      4
3.2           Iris-setosa         5
              Iris-versicolor     3
              Iris-virginica      5
3.3           Iris-setosa         2
              Iris-versicolor     1
              Iris-virginica      3
3.4           Iris-setosa         9
              Iris-versicolor     1
              Iris-virginica      2
3.5           Iris-setosa         6
3.6           Iris-setosa         2
              Iris-virginica      1
3.7           Iris-setosa         3
3.8           Iris-setosa         4
              Iris-virginica      2
3.9           Iris-setosa         2
4.0           Iris-setosa         1
4.1           Iris-setosa         1
4.2           Iris-setosa         1
4.4           Iris-setosa         1
dtype: int64
```

This output indicates that we have two types of combinations.

- Case 1: Gender = F & Gender Group = 1
- Case 2: Gender = M & GenderGroup = 2.

This validates our initial assumption that these two fields essentially portray the same information.

Data Cleaning: handle with missing data

Before getting started to work with your data, it's a good practice to observe it thoroughly to identify missing values and handle them accordingly.

When reading a dataset using Pandas, there is a set of values including 'NA', 'NULL', and 'NaN' that are taken by default to represent a missing value. The full list of default missing value codes is in the '`read_csv`' documentation here. This document also explains how to change the way that '`read_csv`' decides whether a variable's value is missing.

Pandas has functions called `isnull` and `notnull` that can be used to identify where the missing and non-missing values are located in a data frame.

Below we use these functions to count the number of missing and non-missing values in each variable of the datasetr.

In [139…
```
missing= df.isnull().sum()
notMissing= df.notnull().sum()

print("missing:", missing)
```

```
missing: Id               0
SepalLengthCm    0
SepalWidthCm     0
PetalLengthCm    0
PetalWidthCm     0
Species          0
dtype: int64
```

Add and eliminate columns

In some cases it is useful to create or eiminate new columns

In [140…
```
# Add a new column with new data

# Create a column data
NewColumnData = df.SepalWidthCm/df.SepalLengthCm

# Insert that column in the data frame
df.insert(3, "Ratio width to length", NewColumnData, True)

df.head()
```

Out[140…

	Id	SepalLengthCm	SepalWidthCm	Ratio width to length	PetalLengthCm	PetalWidthCm	Species
0	1	5.1	3.5	0.686275	1.4	0.2	Iris-setosa
1	2	4.9	3.0	0.612245	1.4	0.2	Iris-setosa
2	3	4.7	3.2	0.680851	1.3	0.2	Iris-setosa
3	4	4.6	3.1	0.673913	1.5	0.2	Iris-setosa
4	5	5.0	3.6	0.720000	1.4	0.2	Iris-setosa

In [141…

```
# Eliminate inserted column
df.drop("Ratio width to length", axis=1, inplace = True)
df.head()
```

Out[141…

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	1	5.1	3.5	1.4	0.2	Iris-setosa
1	2	4.9	3.0	1.4	0.2	Iris-setosa
2	3	4.7	3.2	1.3	0.2	Iris-setosa
3	4	4.6	3.1	1.5	0.2	Iris-setosa
4	5	5.0	3.6	1.4	0.2	Iris-setosa

In [142…

```
# Add new column derived from existing columns

# The new column is a function of another column
df["SepalLengthMm"] = df["SepalLengthCm"] * 10

df.head()
```

Out[142…

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species	SepalLengthMm
0	1	5.1	3.5	1.4	0.2	Iris-setosa	51.0
1	2	4.9	3.0	1.4	0.2	Iris-setosa	49.0
2	3	4.7	3.2	1.3	0.2	Iris-setosa	47.0
3	4	4.6	3.1	1.5	0.2	Iris-setosa	46.0
4	5	5.0	3.6	1.4	0.2	Iris-setosa	50.0

In [143…

```
# Eliminate inserted column
df.drop("SepalLengthMm", axis=1, inplace = True)
df.head()
```

Out[143…

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	1	5.1	3.5	1.4	0.2	Iris-setosa
1	2	4.9	3.0	1.4	0.2	Iris-setosa
2	3	4.7	3.2	1.3	0.2	Iris-setosa
3	4	4.6	3.1	1.5	0.2	Iris-setosa
4	5	5.0	3.6	1.4	0.2	Iris-setosa

In [144…

```
# Drop several "unused" columns
vars = ["Id"]
df.drop(vars, axis=1, inplace = True)

df.head()
```

Out[144…

	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa

Add and eliminate rows

In some cases it is requiered to add new observations (rows) to the data set

In [145…

```
# Print tail
df.tail()
```

Out[145…

	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
145	6.7	3.0	5.2	2.3	Iris-virginica
146	6.3	2.5	5.0	1.9	Iris-virginica
147	6.5	3.0	5.2	2.0	Iris-virginica
148	6.2	3.4	5.4	2.3	Iris-virginica
149	5.9	3.0	5.1	1.8	Iris-virginica

In [146…

```
## Eliminate inserted row
df.drop([28], inplace = True )
df.tail()
```

Out[146…

	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
145	6.7	3.0	5.2	2.3	Iris-virginica
146	6.3	2.5	5.0	1.9	Iris-virginica
147	6.5	3.0	5.2	2.0	Iris-virginica
148	6.2	3.4	5.4	2.3	Iris-virginica
149	5.9	3.0	5.1	1.8	Iris-virginica

Final remarks

- The understanding of your dataset is essential
    - Number of observations
    - Variables
    - Data types: numerical or categorial
    - What are my variables of interest
- There are several ways to do the same thing
- Cleaning your dataset (dropping out rows with any missing values) is a good practice
- The **Pandas** library provides fancy, high-performance, easy-to-use data structures and data analysis tools

Activity: work with the iris dataset

Repeat this tutorial with the iris data set and respond to the following inquiries

1. Calculate the statistical summary for each quantitative variables. Explain the results
    - Identify the name of each column
    - Identify the type of each column
    - Minimum, maximum, mean, average, median, standar deviation
2. Are there missing data? If so, create a new dataset containing only the rows with the non-missing data
3. Create a new dataset containing only the petal width and length and the type of Flower
4. Create a new dataset containing only the setal width and length and the type of Flower
5. Create a new dataset containing the setal width and length and the type of Flower encoded as a categorical numerical column

In [147…

```
# Download latest version
Ruta = 'Iris.csv'
df = pd.read_csv(Ruta)
df
```

Out[147…

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	1	5.1	3.5	1.4	0.2	Iris-setosa
1	2	4.9	3.0	1.4	0.2	Iris-setosa
2	3	4.7	3.2	1.3	0.2	Iris-setosa
3	4	4.6	3.1	1.5	0.2	Iris-setosa
4	5	5.0	3.6	1.4	0.2	Iris-setosa
...	...	...	...	...	...	...
145	146	6.7	3.0	5.2	2.3	Iris-virginica
146	147	6.3	2.5	5.0	1.9	Iris-virginica
147	148	6.5	3.0	5.2	2.0	Iris-virginica
148	149	6.2	3.4	5.4	2.3	Iris-virginica
149	150	5.9	3.0	5.1	1.8	Iris-virginica

150 rows × 6 columns

In [148…
```
df.describe()
```

Out[148…

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm
count	150.000000	150.000000	150.000000	150.000000	150.000000
mean	75.500000	5.843333	3.054000	3.758667	1.198667
std	43.445368	0.828066	0.433594	1.764420	0.763161
min	1.000000	4.300000	2.000000	1.000000	0.100000
25%	38.250000	5.100000	2.800000	1.600000	0.300000
50%	75.500000	5.800000	3.000000	4.350000	1.300000
75%	112.750000	6.400000	3.300000	5.100000	1.800000
max	150.000000	7.900000	4.400000	6.900000	2.500000

In [149…
```
df.columns
```

Out[149…
```
Index(['Id', 'SepalLengthCm', 'SepalWidthCm', 'PetalLengthCm', 'PetalWidthCm',
       'Species'],
      dtype='object')
```

In [150…
```
df.dtypes
```

Out[150…
```
Id                int64
SepalLengthCm    float64
SepalWidthCm     float64
PetalLengthCm    float64
PetalWidthCm     float64
Species           object
dtype: object
```

In [151…
```
missing= df.isnull().sum()
print("Missing values: ", missing)
```

```
Missing values:  Id               0
SepalLengthCm    0
SepalWidthCm     0
PetalLengthCm    0
PetalWidthCm     0
Species          0
dtype: int64
```

In [152…
```
df1 = df[['PetalLengthCm','PetalWidthCm','Species']]
df1
```

Out[152…

	PetalLengthCm	PetalWidthCm	Species
0	1.4	0.2	Iris-setosa
1	1.4	0.2	Iris-setosa
2	1.3	0.2	Iris-setosa
3	1.5	0.2	Iris-setosa
4	1.4	0.2	Iris-setosa
...	...	...	...
145	5.2	2.3	Iris-virginica
146	5.0	1.9	Iris-virginica
147	5.2	2.0	Iris-virginica
148	5.4	2.3	Iris-virginica
149	5.1	1.8	Iris-virginica

150 rows × 3 columns

In [153…
```
df2 = df[['SepalLengthCm','SepalWidthCm','Species']]
df2
```

Out[153...

	SepalLengthCm	SepalWidthCm	Species
0	5.1	3.5	Iris-setosa
1	4.9	3.0	Iris-setosa
2	4.7	3.2	Iris-setosa
3	4.6	3.1	Iris-setosa
4	5.0	3.6	Iris-setosa
...	...	...	...
145	6.7	3.0	Iris-virginica
146	6.3	2.5	Iris-virginica
147	6.5	3.0	Iris-virginica
148	6.2	3.4	Iris-virginica
149	5.9	3.0	Iris-virginica

150 rows × 3 columns

In [154...

```
categories = df['Species'].unique()
print('Numerical categories', categories)
```

```
Numerical categories ['Iris-setosa' 'Iris-versicolor' 'Iris-virginica']
```

In [155...

```
df['Species Encoded'] = df['Species'].astype('category').cat.codes

df3 = df[['SepalWidthCm','SepalLengthCm','Species Encoded']]
df3
```

Out[155...

	SepalWidthCm	SepalLengthCm	Species Encoded
0	3.5	5.1	0
1	3.0	4.9	0
2	3.2	4.7	0
3	3.1	4.6	0
4	3.6	5.0	0
...	...	...	...
145	3.0	6.7	2
146	2.5	6.3	2
147	3.0	6.5	2
148	3.4	6.2	2
149	3.0	5.9	2

150 rows × 3 columns