

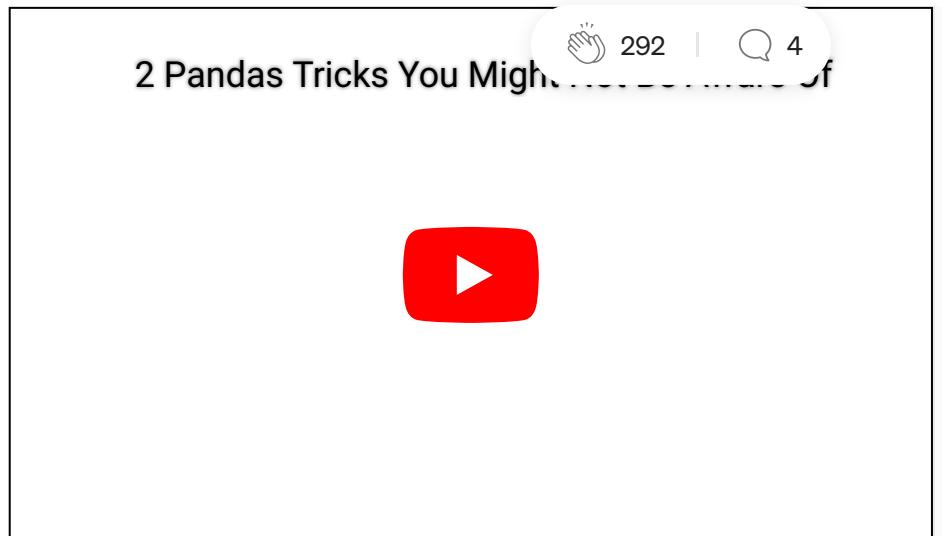
## Motivation

This blog regroups all the Pandas and Python tricks & tips I share on a basis on my [LinkedIn](#) page. I have decided to centralize them into a single blog to help you make the most out of your learning process by easily finding what you are looking for.

The content is divided into two main sections:

- Pandas tricks & tips are related to only Pandas.
- Python tricks & tips related to Python.

If you are more of a video person, you can start watching my series about these tricks on my YouTube channel for more interactivity. Each video covers about two or three tricks at a time.



## Pandas tricks & tips

This section provides a list of all the tricks

### 1. *Create a new column from multiple columns in your dataframe.*

Performing simple arithmetic tasks such as creating a new column as the sum of two other columns can be straightforward.

🤔 But, what if you want to implement a more complex function and use it as the logic behind column creation? Here is where things can get a bit challenging.

Guess what...

✓ **apply** and **lambda** can help you easily apply whatever logic to your columns using the following format:

```
df[new_col] = df.apply(lambda row: func(row), axis=1)
```

where:

- **df** is your dataframe.
- **row** will correspond to each row in your data frame.
- **func** is the function you want to apply to your data frame.
- **axis=1** to apply the function to each row in your data frame.

💡 Below is an illustration.

```

1 import pandas as pd
2
3 # Create the dataframe
4 candidates= {
5     'Name':["Aida","Mamadou","Ismael","Aicha","Fatou", "Khalil"],
6     'Degree':['Master','Master','Bachelor', "PhD", "Master", "PhD"],
7     'From':["Abidjan","Dakar","Bamako", "Abidjan","Konakry", "Lomé"],
8     'Years_exp': [2, 3, 0, 5, 4, 3],
9     'From_office(min)': [120, 95, 75, 80, 100, 34]
10    }
11 candidates_df = pd.DataFrame(candidates)
12
13 """
14 -----My custom function-----
15 """
16 def candidate_info(row):
17
18     # Select columns of interest
19     name = row.Name
20     is_from = row.From
21     year_exp = row.Years_exp
22     degree = row.Degree
23     from_office = row["From_office(min)"]
24
25     # Generate the description from previous variables
26     info = f"""{name} from {is_from} holds a {degree} degree
27             with {year_exp} year(s) experience
28             and lives {from_office} from the office"""
29
30     return info
31
32 """
33 -----Application of the function to the data -----
34 """
35 candidates_df["Description"] = candidates_df.apply(lambda row: candidate_info(row),
axis=1)

```

pandas\_tricks\_.multiple\_cols.py hosted with ❤ by GitHub

[view raw](#)

The `candidate\_info` function combines each candidate's information to create a single description column about that candidate.

Index	Name	Degree	From	Years_exp	From_office(min)	Description
0	Aida	Master	Abidjan	2	120	Aida from Abidjan holds a Master degree with 2 year(s) experience and lives 120 from the office
1	Mamadou	Master	Dakar	3	95	Mamadou from Dakar holds a Master degree with 3 year(s) experience and lives 95 from the office
2	Ismael	Bachelor	Bamako	0	75	Ismael from Bamako holds a Bachelor degree with 0 year(s) experience and lives 75 from the office
3	Aicha	PhD	Abidjan	5	80	Aicha from Abidjan holds a PhD degree with 5 year(s) experience and lives 80 from the office
4	Fatou	Master	Konakry	4	100	Fatou from Konakry holds a Master degree with 4 year(s) experience and lives 100 from the office
5	Khalil	PhD	Lomé	3	34	Khalil from Lomé holds a PhD degree with 3 year(s) experience and lives 34 from the office

Result of Pandas apply and lambda (Image by Author)

## 2. Convert categorical data into numerical ones

This process mainly can occur in the feature engineering phase. Some of its benefits are:

- the identification of outliers, invalid, and missing values in the data.
  - reduction of the chance of overfitting by creating more robust models.
- Use these two functions from Pandas, depending on your need. Examples are provided in the image below.

1 `.cut()` to specifically define your bin edges.

### Scenario

Categorize candidates by expertise with respect to their number of experience, where:

- *Entry level: 0–1 year*
- *Mid-level: 2–3 years*
- *Senior level: 4–5 years*

```

1 seniority = ['Entry level', 'Mid level', 'Senior level']
2 seniority_bins = [0, 1, 3, 5]
3 candidates_df['Seniority'] = pd.cut(candidates_df['Years_exp'],
4                                         bins=seniority_bins,
5                                         labels=seniority,
6                                         include_lowest=True)
7
8 candidates_df

```

	Name	Degree	From	Years_exp	From_office(min)	Seniority
0	Aida	Master	Abidjan	2	120	Mid level
1	Mamadou	Master	Dakar	3	95	Mid level
2	Ismael	Bachelor	Bamako	0	75	Entry level
3	Aicha	PhD	Abidjan	5	80	Senior level
4	Fatou	Master	Konakry	4	100	Senior level
5	Khalil	PhD	Lomé	3	34	Mid level

Result of the .cut function (Image by Author)

2 `.qcut()` to divide your data into equal-sized bins.

It uses the underlying percentiles of the distribution of the data, rather than the edges of the bins.

**Scenario:** categorize the commute time of the candidates into ***good***, ***acceptable***, or ***too long***.

```

1 commute_time_labels = ["good", "acceptable", "too long"]
2 candidates_df["Commute_level"] = pd.qcut(
3                               candidates_df["From_office(min)"],
4                               q = 3,
5                               labels=commute_time_labels
6 )
7 candidates_df

```

qcut\_scenario.py hosted with ❤ by GitHub

[view raw](#)

	Name	Degree	From	Years_exp	From_office(min)	Seniority	Commute_level
0	Aida	Master	Abidjan	2	120	Mid level	too long
1	Mamadou	Master	Dakar	3	95	Mid level	acceptable
2	Ismael	Bachelor	Bamako	0	75	Entry level	good
3	Aicha	PhD	Abidjan	5	80	Senior level	acceptable
4	Fatou	Master	Konakry	4	100	Senior level	too long
5	Khalil	PhD	Lomé	3	34	Mid level	good

Result of the .qcut function (Image by Author)

**Keep in mind** 

- When using `.cut()`: a number of bins = number of labels + 1.
- When using `.qcut()`: a number of bins = number of labels.
- With `.cut()`: set `include_lowest=True`, otherwise, the lowest value will be converted to NaN.

### 3. Select rows from a Pandas Dataframe based on column(s) values

- use `.query()` function by specifying the filter condition.
- the filter expression can contain any operators (<, >, ==, !=, etc.)
- use the @ sign to use a variable in the expression.

```

1 # Get all the candidates with a Master degree
2 ms_candidates = candidates_df.query("Degree == 'Master'")
3
4 # Get non bachelor candidates
5 no_bs_candidates = candidates_df.query("Degree != 'Bachelor'")
6
7 # Get values from list
8 list_locations = ["Abidjan", "Dakar"]
9 candidates = candidates_df.query("From in @list_locations")

```

[filter\\_examples.py](#) hosted with ❤ by GitHub

[view raw](#)

The screenshot shows a Jupyter Notebook cell with the following code:

```

# Import pandas library
import pandas as pd

# Create Dataframe
candidates= {
    'Name':["Aida","Mamadou","Ismael","Aicha","Fatou"],
    'Degree':['Master','Master','Bachelor', 'PhD', "Master"],
    'From':["Abidjan","Dakar","Bamako", "Abidjan","Konakry"]
}
candidates_df = pd.DataFrame(candidates)

# Get all the candidates with a Master degree
ms_candidates = candidates_df.query("Degree == 'Master'")
```

The first execution step shows the original DataFrame:

	Name	Degree	From
0	Aida	Master	Abidjan
1	Mamadou	Master	Dakar
4	Fatou	Master	Konakry

The second execution step shows the filtered DataFrame where Degree is 'Master':

	Name	Degree	From
0	Aida	Master	Abidjan
1	Mamadou	Master	Dakar
4	Fatou	Master	Konakry

The third execution step shows the filtered DataFrame where Degree is not 'Bachelor':

	Name	Degree	From
0	Aida	Master	Abidjan
1	Mamadou	Master	Dakar
3	Aicha	PhD	Abidjan
4	Fatou	Master	Konakry

The fourth execution step shows the filtered DataFrame where From is in the list ['Abidjan', 'Dakar']:

	Name	Degree	From
0	Aida	Master	Abidjan
1	Mamadou	Master	Dakar
3	Aicha	PhD	Abidjan

At the bottom left, there is a Twitter icon followed by the handle '@zoumana\_keita'.

Select rows from a Pandas Dataframe based on column(s) values (Image by Author)

#### 4. Deal with zip files

Sometimes it can be efficient to read and write .zip files without extracting them from your local disk. Below is an illustration.

```
1 import pandas as pd
2
3 """
4 ----- READ ZIP FILES -----
5 """
6 # Case 1: read a single zip file
7 candidate_df_unzip = pd.read_csv('candidates.csv.zip', compression='zip')
8
9 # Case 2: read a file from a folder
10 from zipfile import ZipFile
11
12 # Read the file from a zip folder
13 sales_df = pd.read_csv(ZipFile("data.zip").open('data/sales_df.csv'))
14
15 """
16 """
17 ----- WRITE ZIP FILES -----
18 """
19 # Read data from internet
20 url = "https://raw.githubusercontent.com/keitazoumana/Fastapi-tutorial/master/data/spam.csv"
21 spam_data = pd.read_csv(url, encoding="ISO-8859-1")
22
23 # Save it as a zip file
24 spam_data.to_csv("spam.csv.zip", compression="zip")
25
26 # Check the files sizes
27 from os import path
28 path.getsize('spam.csv') / path.getsize('spam.csv.zip')
```

pandas\_zip\_files.py hosted with ❤ by GitHub

[view raw](#)

#### 5. Select a subset of your Pandas dataframe with specific column types

You can use the `select_dtypes` function. It takes two main parameters: `include` and `exclude`.

- `df.select_dtypes(include = ['type_1', 'type_2', ... 'type_n'])` means I want the subset of my data frame WITH columns of type\_1, type\_2,..., type\_n.

- df.select\_dtypes(exclude = ['type\_1', 'type\_2', ... 'type\_n']) means I want the subset of my data frame WITHOUT columns of type\_1, type\_2,..., type\_n.

💡 Below is an illustration

```

1 # Import pandas library
2 import pandas as pd
3
4 # Read my dataset
5 candidates_df = pd.read_csv("./data/candidates_data.csv")
6
7 # Check the data columns' types
8 candidates_df.dtypes
9
10 # Only select columns of type "object" & "datetime"
11 candidates_df.select_dtypes(include = ["object", "datetime64"])
12
13 # Exclude columns of type "datetime" & "int"
14 candidates_df.select_dtypes(exclude = ["int64", "datetime64"])

```

[select\\_subset\\_column\\_types.py](#) hosted with ❤ by GitHub

[view raw](#)

[select\\_subset\\_column\\_types.py](#)

```

# Import pandas library
import pandas as pd

# Read my dataset
candidates_df = pd.read_csv("./data/candidates_data.csv")

# Check the data columns' types
candidates_df.dtypes

# Only select columns of type "object" & "datetime"
candidates_df.select_dtypes(include = ["object", "datetime64"])

# Exclude columns of type "datetime" & "int"
candidates_df.select_dtypes(exclude = ["int64", "datetime64"])

```

Full_Name	Degree	From	Application_date
Aida, Kone	Master	Abidjan	2022-11-17
Mamadou, Diop	Master	Dakar	2022-09-23
Ismail, Camara	Bachelor	Bamako	2021-12-02
Aicha, Konate	PhD	Abidjan	2022-08-25
Fanta, Koumare	Master	Konakry	2022-01-07
Khalil, Cisse	PhD	Lomé	2022-12-26

Full_Name	Degree	From
Aida, Kone	Master	Abidjan
Mamadou, Diop	Master	Dakar
Ismail, Camara	Bachelor	Bamako
Aicha, Konate	PhD	Abidjan
Fanta, Koumare	Master	Konakry
Khalil, Cisse	PhD	Lomé

Columns subset selection (Image by Author)

## 6. Remove comments from Pandas dataframe column

Imagine that I want clean this data (candidates.csv) by removing comments from the application date column. This can be done on the fly while loading your pandas dataframe using the **comment** parameter as follow:

```
→ clean_data = pd.read_csv(path_to_data, comment='symbol')
```

In my case, `comment='#'` but it could be any other character (|, /, etc.) depending on your case. An illustration is the first scenario.

👉 Wait, what if I want to create a new column for those comments and still remove them from the application date column? An illustration is the second scenario.

```
1 # Import pandas library
2 import pandas as pd
3
4 # Read my messy dataset
5 messy_df = pd.read_csv("./data/candidates_data.csv")
6
7 # FIRST SCENARIO -> REMOVE COMMENTS
8 clean_df = pd.read_csv("./data/candidates_data.csv", comment='#')
9
10 # SECOND SCENARIO -> CREATE NEW COLUMN FOR COMMENTS
11 messy_df[['application_date', 'comment']] =
    messy_df['application_date'].str.split('#', 1, expand=True)
```

pandas\_remove\_comments.py hosted with ❤ by GitHub

[view raw](#)

```
# Import pandas library
import pandas as pd

# Read my messy dataset
messy_df = pd.read_csv("./data/candidates_data.csv")

# FIRST SCENARIO → REMOVE COMMENTS
clean_df = pd.read_csv("./data/candidates_data.csv", comment='#')

# SECOND SCENARIO → CREATE NEW COLUMN FOR COMMENTS
messy_df[['application_date', 'comment']] = messy_df['application_date'].str.split('#', 1, expand=True)
```

Comment symbol					
Full_Name	degree	From	From_office (min)	application_date	comment
0 Aida Kone	Master	Abidjan	120	17/11/2022	# more Interested in Machine Learning
1 Mamadou Diop	Master	Dakar	95	23/09/2022	# open to any type of Data Role
2 Ismael Camara	Bachelor	Bamako	75	02/12/2021	# will be available in 6 months
3 Aicha Konate	PhD	Abidjan	80	25/08/2022	# only interested in Senior positions
4 Fanta Koumare	Master	Konakry	100	07/01/2022	# can relocate to any other cities
5 Khalil Cisse	PhD	Lomé	34	26/12/2022	#wants a minimum monthly salary of ...

Full_Name	degree	From	From_office (min)	application_date
0 Aida Kone	Master	Abidjan	120	17/11/2022
1 Mamadou Diop	Master	Dakar	95	23/09/2022
2 Ismael Camara	Bachelor	Bamako	75	02/12/2021
3 Aicha Konate	PhD	Abidjan	80	25/08/2022
4 Fanta Koumare	Master	Konakry	100	07/01/2022
5 Khalil Cisse	PhD	Lomé	34	26/12/2022

Full_Name	degree	From	From_office (min)	application_date	comment
0 Aida Kone	Master	Abidjan	120	17/11/2022	more Interested in Machine Learning
1 Mamadou Diop	Master	Dakar	95	23/09/2022	open to any type of Data Role
2 Ismael Camara	Bachelor	Bamako	75	02/12/2021	will be available in 6 months
3 Aicha Konate	PhD	Abidjan	80	25/08/2022	only interested in Senior positions
4 Fanta Koumare	Master	Konakry	100	07/01/2022	can relocate to any other cities
5 Khalil Cisse	PhD	Lomé	34	26/12/2022	wants a minimum monthly salary of 3 millions fola

Remove comments from pandas dataframe (Image by Author)



## 7. Print Pandas dataframe in Tabular format from consol

✗ No, the application of the print() function to a pandas data frame does not always render an output that is easy to read, especially for data frames with multiple columns.

✓ If you want to get a nice console-friendly tabular output

Use the .to\_string() function as illustrated below.

```
1 # Import pandas library
2 import pandas as pd
3
4 data_URL = "https://raw.githubusercontent.com/keitazoumana/Experimentation-
  Data/main/vgsales.csv"
5
6 # Read your dataframe
7 video_game_data = pd.read_csv(data_URL)
8
9 """
10 Printing without to_string() function
11 """
12 print(video_game_data.head())
13
14 """
15 Printing with to_string() function
16 """
17 print(video_game_data.head().to_string())
```

pandas\_to\_string.py hosted with ❤ by GitHub

[view raw](#)

```

# Import pandas library
import pandas as pd

data_URL = "https://raw.githubusercontent.com/keitazoumana/Experimentation-Data/main/vgsales.csv"

# Read your dataframe
video_game_data = pd.read_csv(data_URL)

"""
Printing without to_string() function
"""

print(video_game_data.head())

```

Printing with to\_string() function

```

"""
print(video_game_data.head().to_string())

```

Rank	Name	Platform	Year	Genre	Publisher
0	1	Wii Sports	Wii	2006.0	Sports
1	2	Super Mario Bros.	NES	1985.0	Platform
2	3	Mario Kart Wii	Wii	2008.0	Racing
3	4	Wii Sports Resort	Wii	2009.0	Sports
4	5	Pokemon Red/Pokemon Blue	GB	1996.0	Role-Playing

Rank	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
0	41.49	29.02	3.77	8.46	82.74
1	29.08	3.58	6.81	0.77	40.24
2	15.85	12.88	3.79	3.31	35.82
3	15.75	11.01	3.28	2.96	33.00
4	11.27	8.89	10.22	1.00	31.37



Rank	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
0	1	Wii Sports	Wii	2006.0	Sports	41.49	29.02	3.77	8.46	82.74
1	2	Super Mario Bros.	NES	1985.0	Platform	29.08	3.58	6.81	0.77	40.24
2	3	Mario Kart Wii	Wii	2008.0	Racing	15.85	12.88	3.79	3.31	35.82
3	4	Wii Sports Resort	Wii	2009.0	Sports	15.75	11.01	3.28	2.96	33.00
4	5	Pokemon Red/Pokemon Blue	GB	1996.0	Role-Playing	11.27	8.89	10.22	1.00	31.37

@zoumana\_keita\_

## 8. Highlight data points in Pandas

Applying colors to a pandas data frame can be a good way to emphasize certain data points for quick analysis.

✓ This is where pandas.style module comes in handy. It has many features, but is not limited to the followings:

✨ df.style.highlight\_max() to assign a color to the maximum value of each column.

✨ df.style.highlight\_min() to assign a color to the minimum value of each column.

✨ df.style.apply(my\_custom\_function) to apply your custom function to your data frame.

```
1 import pandas as pd
2
3 my_info = {
4     "Salary": [100000.2, 95000.9, 103000.2, 65984.1, 150987.08],
5     "Height": [6.5, 5.2, 5.59, 6.7, 6.92],
6     "weight": [185.23, 105.12, 110.3, 190.12, 200.59]
7 }
8 my_data = pd.DataFrame(my_info)
9
10 """
11 Function to highlight min and max
12 """
13
14 def highlight_min_max(data_frame, min_color, max_color):
15
16     # This first line create a styler object
17     final_data = data_frame.style.highlight_max(color = max_color)
18
19     # On this second line, no need to use ".style"
20     final_data = final_data.highlight_min(color = min_color)
21
22     return final_data
23
24 # Function to apply ORANGE to min and GREEN to max
25 highlight_min_max(my_data, min_color='orange', max_color='green')
26
27 """
28 Custom function: apply RED or GREEN whether data is below or above the mean.
29 """
30
31 def highlight_values(data_row):
32     low_value_color = "background-color:#C4606B ; color: white;"
33     high_value_color = "background-color: #C4DE6B; color: white;"
34     filter = data_row < data_row.mean()
35
36     return [low_value_color if low_value else high_value_color for low_value in filter]
37
38 # Application of my custom function to only 'Height' & 'weight'
39 my_data.style.apply(highlight_values, subset=['Height', 'weight'])
```

```

import pandas as pd

my_info = {
    "Salary": [100000.2, 95000.9, 103000.2, 65984.1, 150987.08],
    "Height": [6.5, 5.2, 5.59, 6.7, 6.92],
    "Weight": [185.25, 105.12, 110.3, 190.12, 200.59]
}
my_data = pd.DataFrame(my_info)

# Function to highlight min and max
# ...

def highlight_min_max(data_frame, min_color, max_color):
    # This first line create a styles object
    final_data = data_frame.style.highlight_max(color = max_color)

    # On this second line, no need to use ".style"
    final_data = final_data.highlight_min(color = min_color)

    return final_data

# Function to update colors to RED and GREEN to mean
highlight_min_max(my_data, min_color='orange', max_color='green')

# Custom function: apply RED or GREEN whether data is below or above the mean.
# ...
def highlight_values(data_row):
    low_value_color = "background-color:#C4E6E8; color: white;"
    high_value_color = "background-color: #C4D6E8; color: white;"
    filter = data_row < data_row.mean()

    return [low_value_color if low_value else high_value_color for low_value in filter]

# Application of my custom function to only 'Height' & 'weight'
my_data.style.apply(highlight_values, subset=['Height', 'weight'])

```

	Salary	Height	weight
0	100000.200000	6.500000	185.230000
1	95000.900000	5.200000	105.120000
2	103000.200000	5.590000	110.300000
3	65984.100000	6.700000	190.120000
4	150987.080000	6.920000	200.590000

	Salary	Height	weight
0	100000.200000	6.500000	185.230000
1	95000.900000	5.200000	105.120000
2	103000.200000	5.590000	110.300000
3	65984.100000	6.700000	190.120000
4	150987.080000	6.920000	200.590000

✓ @zoumana\_keita\_

Highlight data points in Pandas (Image by Author)

## 9. Reduce decimal points in your data

Sometimes, very long decimal values in your data set do not provide significant information and can be painful 😬 to look at.

So, you might want to convert your data to about 2 to 3 decimal points to facilitate your analysis.

- ✓ This is something you can perform using the `pandas.DataFrame.round()` function as illustrated below.

```

1 long_decimals_info = {
2     "Salary": [100000.23400000, 95000.900300, 103000.2300535, 65984.14000450,
3     150987.080345],
4     "Height": [6.501050, 5.270000, 5.5900001050, 6.730001050, 6.92100050],
5     "weight": [185.23000059, 105.1200099, 110.350003, 190.12000000, 200.59000000]
6 }
7
8
9 """
10 Format the data with 2 decimal places
11 """
12 fewer_decimals_df = long_decimals_df.round(decimals=2)
13 fewer_decimals_df

```

[long\\_to\\_few\\_decimals.py](#) hosted with ❤ by GitHub

[view raw](#)



The screenshot shows a Jupyter Notebook cell with the following code:

```

import pandas as pd

long_decimals_info = {
    "Salary": [100000.23400000, 95000.900300, 103000.2300535, 65984.14000450, 150987.080345],
    "Height": [6.501050, 5.270000, 5.5900001050, 6.730001050, 6.92100050],
    "weight": [185.23000059, 105.1200099, 110.350003, 190.12000000, 200.59000000]
}

long_decimals_df = pd.DataFrame(long_decimals_info)

"""
Format the data with 2 decimal places
"""

fewer_decimals_df = long_decimals_df.round(decimals=2)

```

A green box highlights the line `fewer_decimals_df = long_decimals_df.round(decimals=2)`. A green arrow points from this box to the resulting DataFrame output on the right:

	Salary	Height	weight
0	100000.23	6.50	185.23
1	95000.90	5.27	105.12
2	103000.23	5.59	110.35
3	65984.14	6.73	190.12
4	150987.08	6.92	200.59

[@zoumana\\_kella\\_](#)

Reduce decimal points in your data (Image by Author)

## 10. Replace some values in your data frame

You might want to replace some information in your data frame to keep it as up-to-date as possible.

- ✓ This can be achieved using the Pandas `dataframe.replace()` function as illustrated below.

```

1 import pandas as pd
2 import numpy as np
3
4 candidates_info = {
5     'Full_Name':["Aida Kone","Mamadou Diop","Ismael Camara","Aicha Konate",
6                 "Fanta Koumare", "Khalil Cisse"],
7     'degree':['Master', 'MS', 'Bachelor', "PhD", "Masters", np.nan],
8     'From':[np.nan,"Dakar", "Bamako", "Abidjan", "Konakry", "Lomé"],
9     'Age':[23,26,19, np.nan,25, np.nan],
10    }
11
12 candidates_df = pd.DataFrame(candidates_info)
13
14 """
15 Replace Masters, Master by MS
16 """
17 degrees_to_replace = ["Master", "Masters"]
18 candidates_df.replace(to_replace = degrees_to_replace, value = "MS", inplace=True)
19
20 """
21 Replace all the NaN by "Missing"
22 """
23 candidates_df.replace(to_replace=np.nan, value = "Missing", inplace=True)

```

pandas\_replace\_values.py hosted with ❤ by GitHub

[view raw](#)

Full_Name	degree	From	Age
Aida Kone	Master	NaN	23.0
Mamadou Diop	MS	Dakar	26.0
Ismael Camara	Bachelor	Bamako	19.0
Aicha Konate	PhD	Abidjan	NaN
Fanta Koumare	Masters	Konakry	25.0
Khalil Cisse	NaN	Lomé	NaN

Full_Name	degree	From	Age
Aida Kone	MS	NaN	23.0
Mamadou Diop	MS	Dakar	26.0
Ismael Camara	Bachelor	Bamako	19.0
Aicha Konate	PhD	Abidjan	NaN
Fanta Koumare	MS	Konakry	25.0
Khalil Cisse	NaN	Lomé	NaN

Full_Name	degree	From	Age
Aida Kone	MS	Missing	23.0
Mamadou Diop	MS	Dakar	26.0
Ismael Camara	Bachelor	Bamako	19.0
Aicha Konate	PhD	Abidjan	Missing
Fanta Koumare	MS	Konakry	25.0
Khalil Cisse	Missing	Lomé	Missing

import pandas as pd  
import numpy as np  
  
candidates\_info = {  
 'Full\_Name':["Aida Kone","Mamadou Diop","Ismael Camara","Aicha Konate",  
 "Fanta Koumare", "Khalil Cisse"],  
 'degree':['Master', 'MS', 'Bachelor', "PhD", "Masters", np.nan],  
 'From':[np.nan,"Dakar", "Bamako", "Abidjan", "Konakry", "Lomé"],  
 'Age':[23,26,19, np.nan,25, np.nan],  
}  
  
candidates\_df = pd.DataFrame(candidates\_info)  
  
"""
Replace Masters, Master by MS
"""
degrees\_to\_replace = ["Master", "Masters"]
candidates\_df.replace(to\_replace = degrees\_to\_replace, value = "MS", inplace=True)  
  
"""
Replace all the NaN by "Missing"
"""
candidates\_df.replace(to\_replace=np.nan, value = "Missing", inplace=True)

Twitter icon @zoumana\_keita\_

Replace some values in your data frame (Image by Author)

## 11. Compare two data frames and get their differences

Sometimes, when comparing two pandas data frames, not only do you want to know if they are equivalent, but also where the difference lies if they are not equivalent.

✓ This is where the `.compare()` function comes in handy.

★ It generates a data frame showing columns with differences side by side. Its shape is different from  $(0, 0)$  only if the two data being compared are the same.

★ If you want to show values that are equal, set the `keep_equal` parameter to `True`. Otherwise, they are shown as `NaN`.

```
1 import pandas as pd
2 from pandas.testing import assert_frame_equal
3
4 candidates_df = pd.read_csv("data/candidates.csv")
5
6 """
7 Create a second dataframe by changing "Full_Name" & "Age" columns
8 """
9 candidates_df_test = candidates_df.copy()
10 candidates_df_test.loc[0, 'Full_Name'] = 'Aida Traore'
11 candidates_df_test.loc[2, 'Age'] = 28
12
13 """
14 Compare the two dataframes: candidates_df & candidates_df_test
15 """
16 # 1. Comparison showing only unmatching values
17 candidates_df.compare(candidates_df_test)
18
19 # 2. Comparison including similar values
20 candidates_df.compare(candidates_df_test, keep_equal=True)
```

pandas\_compare\_get\_differences.py hosted with ❤ by GitHub

[view raw](#)

```

import pandas as pd
from pandas.testing import assert_frame_equal

candidates_df = pd.read_csv("data/candidates.csv")
"""

Create a second dataframe by changing "Full_Name" & "Age" columns
"""

candidates_df_test = candidates_df.copy()
candidates_df_test.loc[0, 'Full_Name'] = 'Aida Traore'
candidates_df_test.loc[2, 'Age'] = 28

"""

Compare the two dataframes: candidates_df & candidates_df_test
"""

# 1. Comparison showing only unmatching values
candidates_df.compare(candidates_df_test) 1

# 2. Comparison including similar values
candidates_df.compare(candidates_df_test, keep_equal=True) 2

```

Original data

	Full_Name	degree	From	Age
0	Aida Kone	Master	Abidjan	23
1	Mamadou Diop	MS	Dakar	26
2	Ismael Cisse	Bachelor	Bamako	19
3	Aicha Konate	PhD	Abidjan	30
4	Fanta Koumara	Masters	Konakry	25
5	Khalil Cisse	BS	Lomé	24

second data

	Full_Name	degree	From	Age
0	<b>Aida Traore</b>	Master	Abidjan	23
1	Mamadou Diop	MS	Dakar	26
2	<b>Ismael Cisse</b>	Bachelor	Bamako	<b>28</b>
3	Aicha Konate	PhD	Abidjan	30
4	Fanta Koumara	Masters	Konakry	25
5	Khalil Cisse	BS	Lomé	24

Final Results

1

Full_Name		Age		
self	other	self	other	
0	Aida Kone	Aida Traore	NaN	NaN
2	NaN	NaN	19.0	28.0

2

Full_Name		Age		
self	other	self	other	
0	Aida Kone	Aida Traore	23	23
2	Ismael Cisse	Ismael Cisse	19	28

Twitter icon @zoumana\_keita\_

Compare two data frames and get their differences (Image by Author)

## 12. Get a subset of a very large dataset for quick analysis

Sometimes, we just need a subset of a very large dataset for quick analysis. One of the approaches could be to read the whole data in memory before getting your sample.

This can require a lot of memory depending on how big your data is. Also, it can take significant time to read your data.

- ✓ You can use `nrows` parameter in the `pandas read_csv()` function by specifying the number of rows you want.

```
1 # Pandas library
2 import pandas as pd
3
4 # Load execution time
5 %load_ext autotime
6
7 # File to get sample from: Size: 261,6 MB
8 large_data = "diabetes_benchmark_data.csv"
9
10 # Sample size of interest
11 sample_size = 400
12
13 """
14 Approach n°1: Read all the data in memory before getting the sample
15 """
16 read_whole_data = pd.read_csv(large_data)
17 sample_data = read_whole_data.head(sample_size)
18
19 """
20 Approach n°2: Read the sample on the fly
21 """
22 read_sample = pd.read_csv(large_data, nrows=sample_size)
```

select\_subset\_while\_reading.py hosted with ❤ by GitHub

[view raw](#)

```
# Pandas library
import pandas as pd

# Load execution time
%load_ext autotime

# File to get sample from: Size: 261,6 MB
large_data = "diabetes_benchmark_data.csv"

# Sample size of interest
sample_size = 400

"""
Approach n°1: Read all the data in memory before getting the sample
"""

read_whole_data = pd.read_csv(large_data)
sample_data = read_whole_data.head(sample_size) Wall Time ⏳ 4.34 s

"""

Approach n°2: Read the sample on the fly
"""

read_sample = pd.read_csv(large_data, nrows=sample_size) Wall Time ⏳ 12.8ms ✨
```

Twitter icon @zoumana\_keita\_

Get a subset of a very large dataset for quick analysis (Image by Author)

### 13. Transform your data frame from a wide to a long format

Sometimes it can be useful transform your dataframe from a wide to a long format which is more flexible for better analysis, especially when dealing with time series data.

- *What do you mean by wide & long?*

- ✨ Wide format is when you have a lot of columns.
- ✨ Long format on the other side is when you have a lot of rows.

✓ Pandas.melt() is a perfect candidate for this task.

Below is an illustration

```
1 import pandas as pd
2
3 # My experimentation data
4 candidates= {
5     'Name':["Aida","Mamadou","Ismael","Aicha"],
6     'ID': [1, 2, 3, 4],
7     '2017':[85, 87, 89, 91],
8     '2018':[96, 98, 100, 102],
9     '2019':[100, 102, 106, 106],
10    '2020':[89, 95, 98, 100],
11    '2021':[94, 96, 98, 100],
12    '2022':[100, 104, 104, 107],
13        }
14 """
15 Data in wide format
16 """
17 salary_data = pd.DataFrame(candidates)
18
19 """
20 Transformation into the long format
21 """
22 long_format_data = salary_data.melt(id_vars=['Name', 'ID'],
23                                     var_name='Year', value_name='Salary(k$)')
24
```

large\_to\_long.py hosted with ❤ by GitHub

[view raw](#)

import pandas as pd

# My experimentation data

```

candidates = {
    'Name': ["Aida", "Mamadou", "Ismael", "Aicha"],
    'ID': [1, 2, 3, 4],
    '2017': [85, 87, 89, 91],
    '2018': [96, 98, 100, 102],
    '2019': [100, 102, 106, 106],
    '2020': [89, 95, 98, 100],
    '2021': [94, 96, 98, 100],
    '2022': [100, 104, 104, 107],
}

```

Data in wide format ①

```

salary_data = pd.DataFrame(candidates)

```

Transformation into the long format ②

```

long_format_data = salary_data.melt(id_vars=['Name', 'ID'],
                                     var_name='Year',
                                     value_name='Salary(k$)')

```

① Wide format

Name	ID	2017	2018	2019	2020	2021	2022
Aida	1	85	96	100	89	94	100
Mamadou	2	87	98	102	95	96	104
Ismael	3	89	100	106	98	98	104
Aicha	4	91	102	106	100	100	107

② Long format ⭐⭐

Name	ID	Year	Salary(k\$)
Aida	1	2017	85
Mamadou	2	2017	87
Ismael	3	2017	89
Aicha	4	2017	91
Aida	1	2018	96
Mamadou	2	2018	98
Ismael	3	2018	100
Aicha	4	2018	102
Aida	1	2019	100
Mamadou	2	2019	102
Ismael	3	2019	106
Aicha	4	2019	106
Aida	1	2020	89
Mamadou	2	2020	95
Ismael	3	2020	98
Aicha	4	2020	100
Aida	1	2021	94
Mamadou	2	2021	96
Ismael	3	2021	98
Aicha	4	2021	100
Aida	1	2022	100
Mamadou	2	2022	104
Ismael	3	2022	104
Aicha	4	2022	107

@zoumana\_keita\_

Transform your data frame from a wide to a long format (Image by Author)

## 14. Reduce the size of your Pandas data frame by ignoring the index

Do you know that you can reduce the size of your Pandas data frame by ignoring the index when saving it?

- Something like `index=False` when saving the file.

Below is an illustration.

```
1 import pandas as pd
2
3 # Read data from Github
4 URL = "https://raw.githubusercontent.com/keitazoumana/Experimentation-
Data/main/diabetes.csv"
5 data = pd.read_csv(URL)
6
7 # Create large data by repeating each row 10000 times
8 large_data = data.loc[data.index.repeat(10000)]
9
10 #####
11 SAVE WITH INDEX
12 #####
13 large_data.to_csv("large_data_with_index.csv")
14
15 # Check the size of the file
16 !ls -GFlash large_data_with_index.csv
17
18 #####
19 SAVE WITHOUT INDEX
20 #####
21 large_data.to_csv("large_data_without_index.csv", index = False)
22
23 # Check the size of the file
24 !ls -GFlash large_data_without_index.csv
```

ignore\_index.py hosted with ❤ by GitHub

[view raw](#)

```
import pandas as pd

# Read data from Github
URL = "https://raw.githubusercontent.com/keitazoumana/Experimentation-Data/main/diabetes.csv"
data = pd.read_csv(URL)

# Create large data by repeating each row 10000 times
large_data = data.loc[data.index.repeat(10000)]

"""
SAVE WITH INDEX
"""
large_data.to_csv("large_data_with_index.csv")

# Check the size of the file
!ls -GFlash large_data_with_index.csv → File Size
250M

"""
SAVE WITHOUT INDEX
"""
large_data.to_csv("large_data_without_index.csv", index = False) → Decreased
# Check the size of the file
!ls -GFlash large_data_without_index.csv → File Size
222M
by 28 Megabytes
```

Decreased  
by 28 Megabytes

Reduce the size of your Pandas data frame by ignoring the index (Image by Author)

## 15. Parquet instead of CSV

Very often, I don't manually look at the content of a CSV or Excel file that will be used by Pandas for further analysis.

If that's your case, maybe you should not use .CSV anymore and think of a better option.

Especially if you are only concerned about

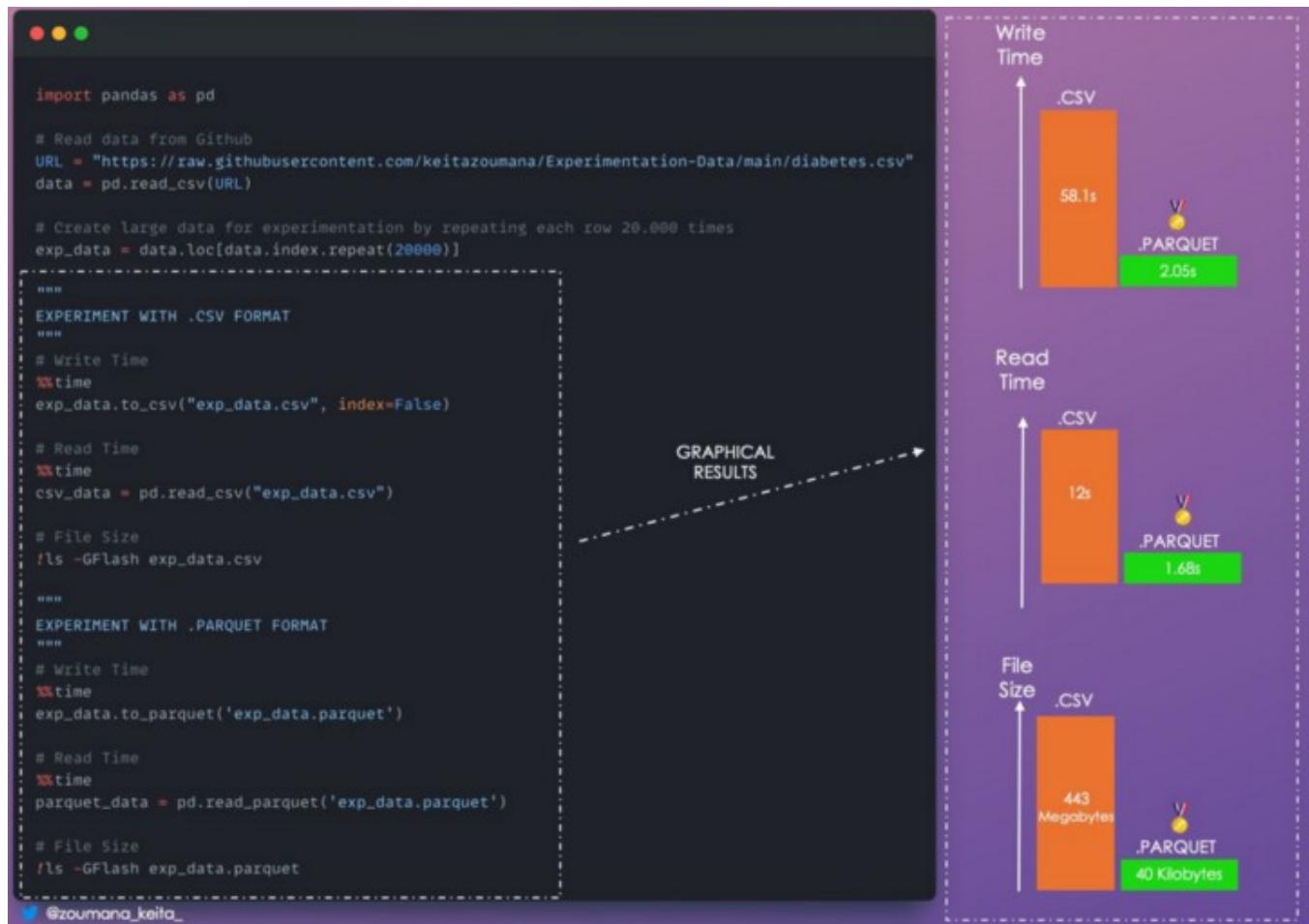
- ✨ Processing speed
- ✨ Speed in saving and loading
- ✨ Disk space occupied by the data frame

In that case, **.parquet** format is your best option as illustrated below.

```
1 import pandas as pd
2
3 # Read data from Github
4 URL = "https://raw.githubusercontent.com/keitazoumana/Experimentation-
Data/main/diabetes.csv"
5 data = pd.read_csv(URL)
6
7 # Create large data for experimentation by repeating each row 20.000 times
8 exp_data = data.loc[data.index.repeat(20000)]
9
10 #####
11 EXPERIMENT WITH .CSV FORMAT
12 #####
13 # Write Time
14 %%time
15 exp_data.to_csv("exp_data.csv", index=False)
16
17 # Read Time
18 %%time
19 csv_data = pd.read_csv("exp_data.csv")
20
21 # File Size
22 !ls -GFlash exp_data.csv
23
24 #####
25 EXPERIMENT WITH .PARQUET FORMAT
26 #####
27 # Write Time
28 %%time
29 exp_data.to_parquet('exp_data.parquet')
30
31 # Read Time
32 %%time
33 parquet_data = pd.read_parquet('exp_data.parquet')
34
35 # File Size
36 !ls -GFlash exp_data.parquet
```

csv\_vs\_parquet.py hosted with ❤ by GitHub

[view raw](#)



Parquet instead of CSV (Image by Author)

## 16. Transform your data frame into a markdown

It is always better to print your data frame in a way that makes it easier to understand.

- ✓ One way of doing that is to render it in a markdown format using the `.to_markdown()` function.

💡 Below is an illustration

```

import pandas as pd

# Data URL
data_URL = "https://raw.githubusercontent.com/keitazoumana/Experimentation-Data/main/vgsales.csv"

# Read as a dataframe
video_game_data = pd.read_csv(data_URL)

# Get the first 5 rows
head_df = video_game_data.head()

# Printing your dataframe
print(head_df)

```

Without Markdown

Rank	Name	Platform	Year	Genre	Publisher	
0	1	Wii Sports	Wii	2006.0	Sports	Nintendo
1	2	Super Mario Bros.	NES	1985.0	Platform	Nintendo
2	3	Mario Kart Wii	Wii	2008.0	Racing	Nintendo
3	4	Wii Sports Resort	Wii	2009.0	Sports	Nintendo
4	5	Pokemon Red/Pokemon Blue	GB	1996.0	Role-Playing	Nintendo

NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
41.49	29.02	3.77	8.46	82.74
29.08	3.58	6.81	0.77	40.24
15.85	12.88	3.79	3.31	35.82
15.75	11.01	3.28	2.96	33.00
11.27	8.89	10.22	1.00	31.37

```

# Print the dataframe in a grid format.
print(head_df.to_markdown(tablefmt='grid'))

```

With Markdown

Rank	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales	
0	1	Wii Sports	Wii	2006	Sports	Nintendo	41.49	29.02	3.77	8.46	82.74
1	2	Super Mario Bros.	NES	1985	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24
2	3	Mario Kart Wii	Wii	2008	Racing	Nintendo	15.85	12.88	3.79	3.31	35.82
3	4	Wii Sports Resort	Wii	2009	Sports	Nintendo	15.75	11.01	3.28	2.96	33.00
4	5	Pokemon Red/Pokemon Blue	GB	1996	Role-Playing	Nintendo	11.27	8.89	10.22	1	31.37

@zoumania\_keita\_

## 17. Format Date Time column

When loading Pandas dataframes, date columns are represented as **object** by default, which is not ✗ the correct date format.

- ✓ You can specify the target column in the **parse\_dates** argument to get the correct column type.

The diagram illustrates the process of parsing a date column in a pandas DataFrame. It starts with a terminal window showing the initial state of the DataFrame, where the 'Application\_date' column is of type 'object'. A callout bubble labeled 'Before parsing' points to this row. An orange arrow points from this state down to a code editor window. The code editor shows Python code for reading a CSV file and displaying its information. In the third code block, there is a line of code: `parse_dates = ["Application_date"]`, which is highlighted with a green box. Another orange arrow points from this code line down to a second terminal window at the bottom. This second terminal window shows the result of the parsing: the 'Application\_date' column has been converted to type 'datetime64[ns]'. A green callout bubble labeled 'After parsing' points to this row.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6 entries, 0 to 5
Data columns (total 3 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   Name            6 non-null      object  
 1   Degreee         6 non-null      object  
 2   Application_date 6 non-null     object 

import pandas as pd

# Read the dataframe
candidates_df = pd.DataFrame("candidates.csv")

# Show columns information
candidates_df.info()

# Use the "parse_date" attribute
candidates_df = pd.DataFrame("candidates.csv",
                             parse_dates = ["Application_date"])

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6 entries, 0 to 5
Data columns (total 3 columns):
 #   Column           Non-Null Count  Dtype    
--- 
 0   Name            6 non-null      object    
 1   Degreee         6 non-null      object    
 2   Application_date 6 non-null     datetime64[ns]
```

🐦 @zoumana\_keita\_

DateTime Formating

## Python tips and tricks

### 1. Create a progress bar with tqdm and rich

Using the progress bar is beneficial when you want to have a visual status of a given task.

```
#!pip -q install rich
from rich.progress import track
from tqdm import tqdm
import time
```

## Implement the callback function

```
def compute_double(x):
    return 2*x
```

## Create the progress bars

```
1 final_dict_doubles = {}
2
3 for i in track(range(20), description="Computing 2.n..."):
4     final_dict_doubles[f"Value = {i}"] = f"double = {compute_double(i)}"
5
6     # Sleep the process to highlight the progress
7     time.sleep(0.8)
```

rich\_progress\_bar.py hosted with ❤ by GitHub

[view raw](#)

rich progress bar implementation

```
1 for i in tqdm(range(20), desc="Computing 2.n..."):
2     final_dict_doubles[f"Value = {i}"] = f"double = {compute_double(i)}"
3
4     # Sleep the process to highlight the progress
5     time.sleep(1)
```

tqdm\_progress\_bar.py hosted with ❤ by GitHub

[view raw](#)

tqdm progress bar implementation

## 2. Get day, month, year, day of the week, the month of the year

```

1 candidates= {
2     'Name':["Aida","Mamadou","Ismael","Aicha","Fatou", "Khalil"],
3     'Degree':['Master','Master','Bachelor', "PhD", "Master", "PhD"],
4     'From':["Abidjan","Dakar","Bamako", "Abidjan","Konakry", "Lomé"],
5     'Application_date': ['11/17/2022', '09/23/2022', '12/2/2021',
6                           '08/25/2022', '01/07/2022', '12/26/2022']
7 }
8 candidates_df = pd.DataFrame(candidates)
9 candidates_df['Application_date'] = pd.to_datetime(candidates_df["Application_date"])
10
11 # GET the Values
12 application_date = candidates_df["Application_date"]
13
14 candidates_df["Day"] = application_date.dt.day
15 candidates_df["Month"] = application_date.dt.month
16 candidates_df["Year"] = application_date.dt.year
17 candidates_df["Day_of_week"] = application_date.dt.day_name()
18 candidates_df["Month_of_year"] = application_date.dt.month_name()

```

use\_of\_dt\_accessor.py hosted with ❤ by GitHub

[view raw](#)

	Name	Degree	From	Application_date
0	Aida	Master	Abidjan	2022-11-17
1	Mamadou	Master	Dakar	2022-09-23
2	Ismael	Bachelor	Bamako	2021-12-02
3	Aicha	PhD	Abidjan	2022-08-25
4	Fatou	Master	Konakry	2022-01-07
5	Khalil	PhD	Lomé	2022-12-26

	Name	Degree	From	Application_date	Day	Month	Year	Day_of_week	Month_of_year
0	Aida	Master	Abidjan	2022-11-17	17	11	2022	Thursday	November
1	Mamadou	Master	Dakar	2022-09-23	23	9	2022	Friday	September
2	Ismael	Bachelor	Bamako	2021-12-02	2	12	2021	Thursday	December
3	Aicha	PhD	Abidjan	2022-08-25	25	8	2022	Thursday	August
4	Fatou	Master	Konakry	2022-01-07	7	1	2022	Friday	January
5	Khalil	PhD	Lomé	2022-12-26	26	12	2022	Monday	December

Get day, month, year, day of the week, the month of the year (Image by author)

### 3. Smallest and largest values of a column

If you want to get the rows with the largest or lowest values for a given column, you can use the following functions:

✨ df.nlargest(N, “Col\_Name”) → top N rows based on Col\_Name

✨ df.nsmallest(N, "Col\_Name") → N smallest rows based on Col\_Name

✨ Col\_Name is the name of the column you are interested in.

The screenshot shows a Jupyter Notebook interface. On the left, there are two code cells. The top cell contains:import pandas as pd  
candidates\_df = pd.DataFrame("candidates.csv")  
  
# 3 Youngest Candidates  
candidates\_df.nsmallest(3, "Age")The bottom cell contains:# 3 Oldest Candidates  
candidates\_df.nlargest(3, "Age")On the right, there are two tables representing the results of these operations. The first table, titled "3 Youngest Candidates", shows three rows: Ismael Cisse (Bachelor, Bamako, 19), Aida Kone (Master, Abidjan, 23), and Khalil Cisse (BS, Lomé, 24). The second table, titled "3 Oldest Candidates", shows three rows: Aicha Konate (PhD, Abidjan, 30), Mamadou Diop (MS, Dakar, 26), and Fanta Koumara (Masters, Konakry, 25).

Smallest and largest values illustration (Image by Author)

#### 4. Ignore the log output of the pip install command

Sometimes when installing a library from your jupyter notebook, you might not want to have all the details about the installation process generated by the default pip install command.

✓ You can specify the -q or – quiet option to get rid of that information.

Below is an illustration

The screenshot shows a terminal window with two side-by-side panes. The left pane is titled "Install Without '-q' option" and contains the command: !pip install spacy-transformers. The right pane is titled "Install With '-q' option" and contains the command: !pip -q install spacy-transformers. The terminal window has a dark background and a light-colored text area. At the bottom, it says "Zoumana KEITA".

pip install illustration (Animation by Author)

#### 5. Run multiple commands in a single notebook cell

The exclamation mark ‘!’ is essential to successfully run a shell command from your Jupyter notebook.

However, this approach can be quite repetitive  when dealing with multiple commands or a very long and complicated one.

- ✓ A better way to tackle this issue is to use the `%%bash` expression at the beginning of your notebook cell.

 Below is an illustration

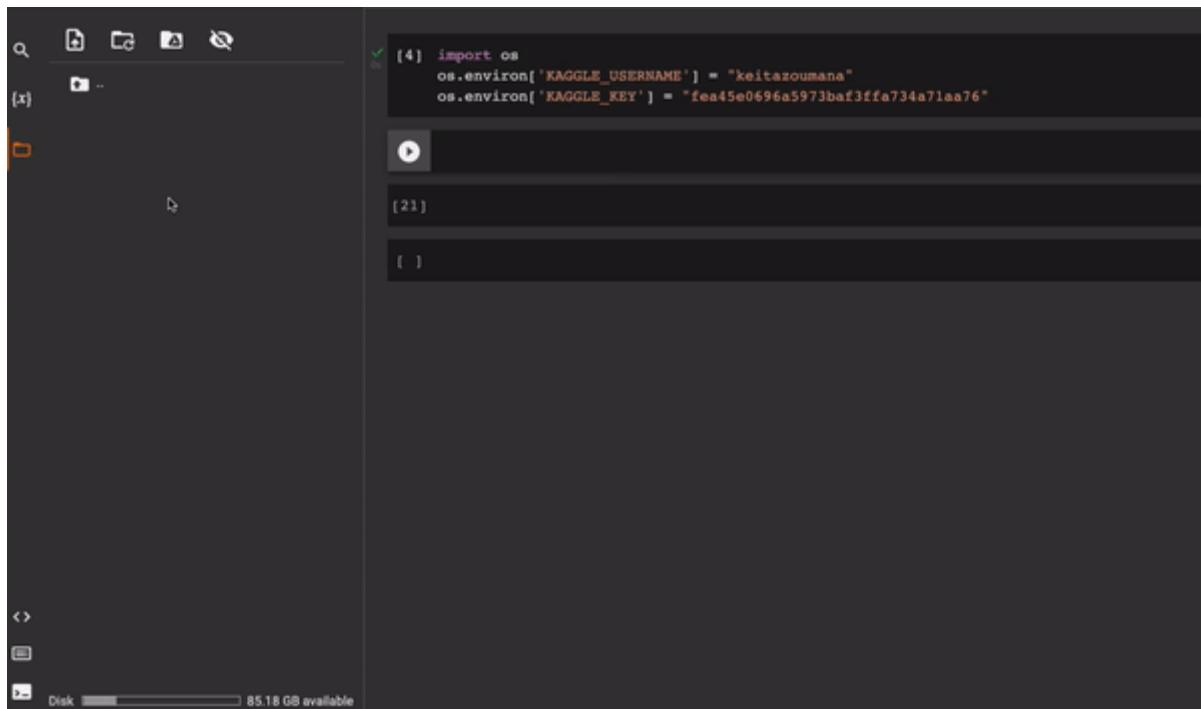


Illustration of `%%bash` statement (Animation by Autor)

## 6. Virtual environment.

A Data Science project can involve multiple dependencies, and dealing with all of them can be a bit annoying. 

✨ A good practice is to organize your project in a way that it can be easily shared with your team members and reproduced with the least amount of effort.

- ✓ One way of doing this is to use virtual environments.

### Create virtual environment and install libraries.

→ Install the virtual environment module.

```
pip install virtualenv
```

→ Create your environment by giving a meaningful name.

```
virtualenv [your_environment_name]
```

→ Activate your environment.

```
source [your_environment_name]/bin/activate
```

→ Start installing the dependencies for your project.

```
pip install pandas
```

...

All this is great 🙌, BUT... the virtual environment you just created is local to your machine 😞.

### What to do? 🤔

💡 You need to permanently save those dependencies in order to share them with others using this command:

→ pip freeze > requirements.txt

This will create requirements.txt file containing your project dependencies.

◀ Finally, anyone can install the exact same dependencies by running this command:

```
→ pip install -r requirements.txt
```

## 7. Run multiple metrics at once

Scikit learn metrics

```
1  """
2  Individual imports
3  """
4  from sklearn.metrics import precision_score, recall_score, f1_score
5
6  y_true = [0, 1, 2, 0, 1, 2]
7  y_pred = [0, 2, 1, 0, 0, 1]
8
9  print("Precision: ", precision_score(y_true, y_pred, average='macro'))
10 print("Recall: ", recall_score(y_true, y_pred, average='macro'))
11 print("F1 Score: ", f1_score(y_true, y_pred, average='macro'))
12
13
14 """
15 Single Line import
16 """
17 from sklearn.metrics import precision_recall_fscore_support
18
19 precision, recall, f1_score, _ = precision_recall_fscore_support(y_true,
20                     y_pred,
21                     average='macro')
22 print(f"Precision: {precision}")
23 print(f"Recall: {recall}")
24 print(f"F1 Score: {f1_score}")
```

multiple\_metrics.py hosted with ❤ by GitHub

[view raw](#)

## 8. Chain multiple lists as a single sequence

You can use a single for loop to iterate through multiple lists as a single sequence .

 This can be achieved using the `chain()` function from Python `itertools` module.

```
from itertools import chain

# List of candidates
senegal = ['Sadio', 'Aicha', 'Abbul Kadr']
cote_dIvoire = ['Serge', 'Moussa', 'Nabintou']

# Iterate through the two lists at the same time
for candidate in chain(senegal, cote_dIvoire):
    print(candidate)
```

Chain lists

```
Sadio
Aicha
Abbul Kadr
Serge
Moussa
Nabintou
```

Final Result

造福 @zoumana\_keita\_

List chaining

## 9. Pretty print of JSON data

? Have ever wanted to print your JSON data in a correct indented format for better visualization?

✓ The indent parameter of the dumps() method can be used to specify the indentation level of your formatted string output.

A screenshot of a terminal window on a Mac OS X system. The window title bar shows the standard red, yellow, and green close/minimize/maximize buttons. The terminal itself contains the following Python code:

```
# Import the library
import json

# Initialize the JSON data
candidates_data = '[{"Name": "Abdoul", "Age": 23, \
    "Degrees": ["Bachelor", "Master"]}, \
    {"Name": "Mariam", "Age": 28, \
    "Degrees": ["Bachelor", "Master", "PhD"]}]'

# Create Python object from JSON string
candidates_Json = json.loads(candidates_data)

# Print with indentation level of 4
print(json.dumps(candidates_Json, indent = 4))
```

The code uses the `json` module to parse a JSON string into a Python list of dictionaries. It then prints the list back out using `json.dumps` with an indentation level of 4. A callout bubble points to the `indent = 4` argument with the text "Create indentation level of 4". To the right of the terminal, a box displays the resulting pretty-printed JSON output:

```
[ {
    "Name": "Abdoul",
    "Age": 23,
    "Degrees": [
        "Bachelor",
        "Master"
    ]
},
{
    "Name": "Mariam",
    "Age": 28,
    "Degrees": [
        "Bachelor",
        "Master",
        "PhD"
    ]
}]
```

Pretty print your JSON data

## 10. Unit testing

Do you Test Your Code? ✎

I mean do you perform Unit Testing?

No matter if you are Data Scientist or a Software Developer, Unit testing is an important step to make sure the features being implemented meet the expected behavior.

This is undoubtedly beneficial on many levels:

- ✨ Better quality 💎 code.
  - ✨ Allows simpler and more agile code when adding new features.
  - ✨ Reduces cost 💰 by saving dev time ⏳ and avoiding later stages of error discovery.
  - ✨ Much More ...
- ✓ With `unittest`, you can perform unit testing like a pro 😎

Below is an illustration 💡

```
# Import the library
import unittest

# Function to test
def product(a, b):
    return a * b

class TestProduct(unittest.TestCase):

    # Implement the unit test
    def test_product_two_variables(self):

        # Is 16 not equal to product(4, 3)
        self.assertNotEqual(16, product(4, 3))

Ran 1 test in 0.001s
OK
<unittest.main.TestProgram at 0x10c331f00>
```

Is 16 different from  $4 \times 3$  ?



@zoumana\_keita\_

Unit test illustration

## 11. Iterate over multiple lists

Iterating over multiple lists simultaneously can be beneficial when trying to map 📈 information from those lists.

- ✓ My go-to approach is the Python `zip` function.

Below is an illustration 💡

```
# List of names  
names = ['Sadio', 'Aicha', 'Nabintou', 'Ahmed']  
  
# List of locations  
locations = ['Dakar', 'Bamako', 'Abidjan', 'Conakry']
```

```
# Simultaneous iteration
```

```
for name, location in zip(names, locations):  
    print(f'{name}: {location}')
```

Iteration  
through the two  
lists

```
Sadio: Dakar  
Aicha: Bamako  
Nabintou: Abidjan  
Ahmed: Conakry
```

🐦 @zoumana\_keita\_

Iterate over multiple lists

## 12. Alternative to nested for loops

Raise your hand if you have once used nested loops 🤝🏿

This is most of the time inevitable when a program gets complicated.

However, using nested loops get makes your program harder to read 🧐 and maintain 🤷.

- ✓ You can use the Python built-in `product()` function instead.

Below is an illustration 💡

```
# Define my lists
first_list = [4, 12, 6]
second_list = [5, 2, 19]
third_list = [7, 2, 3]

"""
Run the cartesian product
"""

# Method 1
for fst in first_list:
    for snd in second_list:
        for trd in third_list:
            if(fst * snd * trd == 24):
                print(f' 1st: {fst} 2nd: {snd} 3rd: {trd}')


# Import the product module from itertools
from itertools import product

for fst, snd, trd in product(first_list, second_list, third_list):
    if(fst * snd * trd == 24):
        print(f' 1st: {fst} 2nd: {snd} 3rd: {trd}')


X
```

3-level nested  
for loop

```
1st: 4 2nd: 2 3rd: 3
1st: 6 2nd: 2 3rd: 2
```

✓

Neater and cleaner ✨

🐦 @zoumana\_keita\_

Solution to nested for loops

### 13. Text preprocessing made easy

Text STX ETX preprocessing has never been easy.

? How many functions or regular expressions do you have to write to perform basic text preprocessing tasks like:

✨ Fixing Unicode

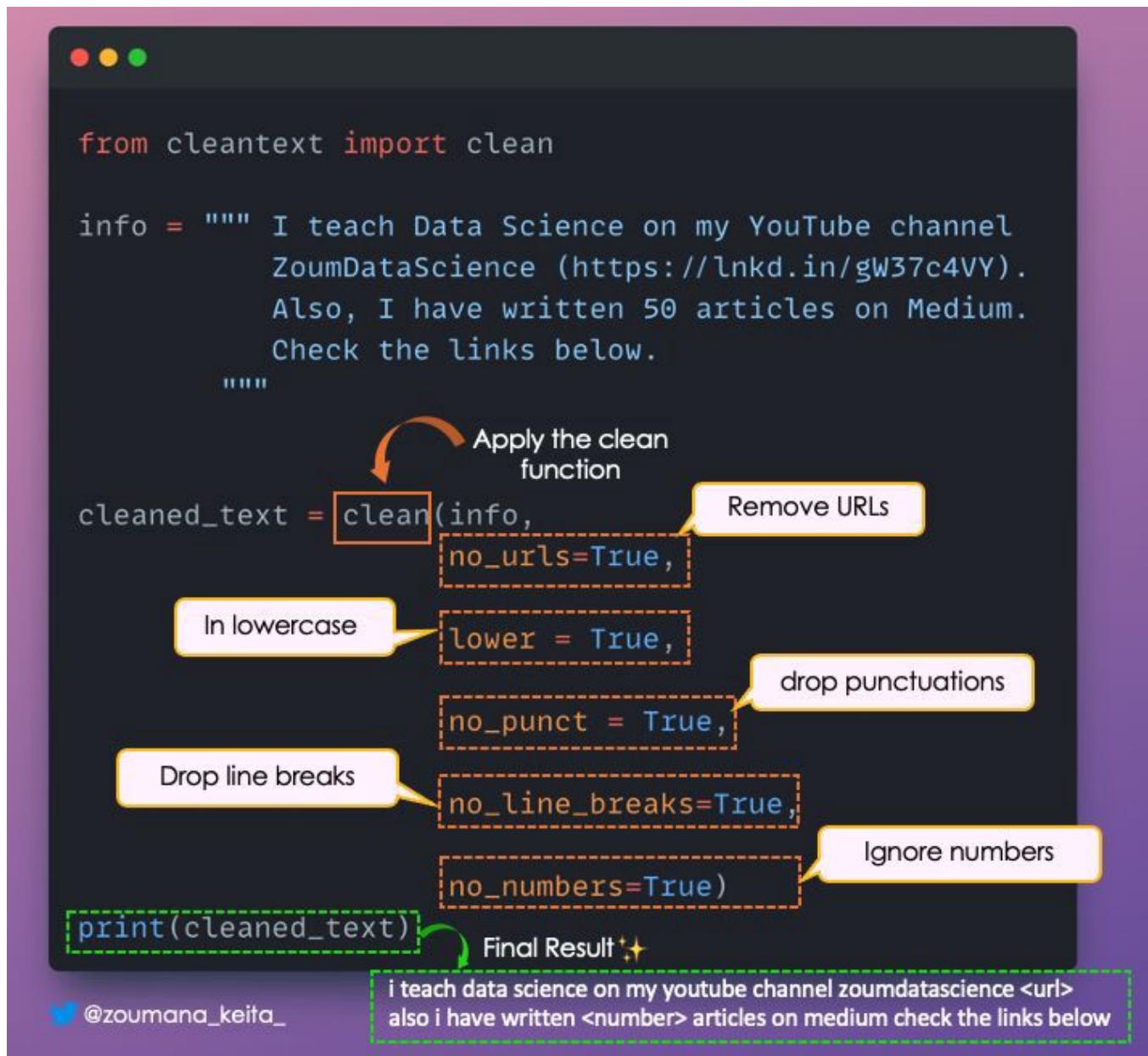
✨ Removing URLs

✨ Getting rid of digits, punctuation, etc?

Those tasks are not only time-consuming ⏳ but also may increase in complexity ↗ depending on the text.

- ✓ Using the `cleantext` Python library can take away all that burden.

Below is an illustration 



```
from cleantext import clean

info = """ I teach Data Science on my YouTube channel
ZoumDataScience (https://lnkd.in/gW37c4VY).
Also, I have written 50 articles on Medium.
Check the links below.
"""

cleaned_text = clean(info,
                     no_urls=True,
                     lower=True,
                     no_punct=True,
                     no_line_breaks=True,
                     no_numbers=True)

print(cleaned_text)
```

Final Result ✨

@zoumana\_keita\_

i teach data science on my youtube channel zoumdatascience <url>  
also i have written <number> articles on medium check the links below

Text preprocessing illustration (Image by Author)

## Conclusion

Thank you for reading! 

I hope you found this list of Python and Pandas tricks helpful! Keep an eye on here, because the content will be maintained with more tricks on a daily basis.

Also, If you like reading my stories and wish to support my writing, consider [becoming a Medium member](#). With a \$ 5-a-month commitment, you unlock unlimited access to stories on Medium.