In [1]:
```
import pandas as pd
import numpy as np
```

In [2]:
```
df = pd.read_excel(r"C:\DOCS\trabajo_analista\RankinUniversidadesDelMundo.xlsx")
```

In [25]:
```
df
```

Out[25]:

	Rank	University name	locationLocation	Number of Studnet	Number of student per staffs	International Student	Female : male ratio	Universidad	Ciudad
0	1	University of Oxford	United Kingdom	20.965	10.6	0.42	1900-01-02 00:52:00	University of Oxford	None
1	2	Harvard University	United States	21.887	9.6	0.25	1900-01-02 02:50:00	Harvard University	None
2	3	University of Cambridge	United Kingdom	20.185	11.3	0.39	1900-01-01 23:53:00	University of Cambridge	None
3	3	Stanford University	United States	16.164	7.1	0.24	1900-01-01 22:54:00	Stanford University	None
4	5	Massachusetts Institute of Technology	United States	11.415	8.2	0.33	40 : 60	Massachusetts Institute of Technology	None
...	...	...	...	...	...	...	...	...	...
2340	Reporter	York St John University	United Kingdom	6.315	18.6	0.12	1900-01-02 17:35:00	York St John University	None
2341	Reporter	Yusuf Maitama Sule University, Kano	Nigeria	12.880	33.0	0	1900-01-02 00:52:00	Yusuf Maitama Sule University	Kano
2342	Reporter	Zhytomyr Polytechnic State University	Ukraine	3.869	15.4	0.01	34 : 66	Zhytomyr Polytechnic State University	None
2343	Reporter	Ziauddin University	Pakistan	4.906	8.8	0.01	1900-01-02 15:37:00	Ziauddin University	None
2344	Reporter	Zarqa University	Jordan	5.768	18.1	0.32	1900-01-01 23:53:00	Zarqa University	None

2345 rows × 9 columns

Eliminar Duplicados

In [4]:
```
df.drop_duplicates()
```

Out[4]:

	Rank	University name	locationLocation	Number of Studnet	Number of student per staffs	International Student	Female : male ratio
0	1	University of Oxford	United Kingdom	20.965	10.6	0.42	1900-01-02 00:52:00
1	2	Harvard University	United States	21.887	9.6	0.25	1900-01-02 02:50:00
2	3	University of Cambridge	United Kingdom	20.185	11.3	0.39	1900-01-01 23:53:00
3	3	Stanford University	United States	16.164	7.1	0.24	1900-01-01 22:54:00
4	5	Massachusetts Institute of Technology	United States	11.415	8.2	0.33	40 : 60
...	...	...	...	...	...	...	...
2340	Reporter	York St John University	United Kingdom	6.315	18.6	0.12	1900-01-02 17:35:00
2341	Reporter	Yusuf Maitama Sule University, Kano	Nigeria	12.880	33.0	0	1900-01-02 00:52:00
2342	Reporter	Zhytomyr Polytechnic State University	Ukraine	3.869	15.4	0.01	34 : 66
2343	Reporter	Ziauddin University	Pakistan	4.906	8.8	0.01	1900-01-02 15:37:00
2344	Reporter	Zarqa University	Jordan	5.768	18.1	0.32	1900-01-01 23:53:00

2345 rows × 7 columns

Eliminar cualquier columa

In [ ]:
```
df.drop(columns = "El nombre de la columna que deseas eliminar")
```

Eliminar todos los datos diferentes que no esten en el abecedario o en los numeros del 1 al 9

In [5]:
```
df["University name"] = df["University name"].str.replace('[^a-zA-Z0-9]',' ')
```

```
C:\Users\ASUS\AppData\Local\Temp\ipykernel_20300\1698705659.py:1: FutureWarning: The default value of regex wil
l change from True to False in a future version.
  df["University name"] = df["University name"].str.replace('[^a-zA-Z0-9]',' ')
```

In [6]:
```
print(df["University name"])
```

```
0                        University of Oxford
1                          Harvard University
2                       University of Cambridge
3                           Stanford University
4       Massachusetts Institute of Technology
                          ...
2340                   York St John University
2341     Yusuf Maitama Sule University  Kano
2342    Zhytomyr Polytechnic State University
2343                       Ziauddin University
2344                          Zarqa University
Name: University name, Length: 2345, dtype: object
```

Configurar las opciones de pandas para mostrar todos los datos

In [28]:
```
pd.set_option('display.max_rows', None)
```

Restablecer la configuracion para que no muestre todos los datos

In [36]:
```
pd.reset_option('display.max_rows')
```

Muchas veces no podemos aplicar codigo en datos string porque no lo son. Para convertirlos en stream hay que hacer los siguiente:

In [7]:
```
df["Number of Studnet"] = df["Number of Studnet"].apply(lambda x: str(x))
```

In [8]:
```
df["Number of Studnet"]
```

Out[8]:
```
0       20.965
1       21.887
2       20.185
3       16.164
4       11.415
         ...
2340     6.315
2341     12.88
2342     3.869
2343     4.906
2344     5.768
Name: Number of Studnet, Length: 2345, dtype: object
```

Para remplazar los datos que no quieres por un espacio en blanco se codea lo sgte

In [15]:
```
df["Female : male ratio"] = df["Female : male ratio"].str.replace('n/a','')
```

Separar columnas pegadas en diferentes columnas, por ejemplo en DIRECCION colocan la calle/cra, la ciudad y el estados o país

In [11]:
```
df[["Universidad", "Ciudad"]] = df["University name"].str.split(',',1, expand=True)
```

```
C:\Users\ASUS\AppData\Local\Temp\ipykernel_19796\4066816243.py:1: FutureWarning: In a future version of pandas
all arguments of StringMethods.split except for the argument 'pat' will be keyword-only.
  df[["Universidad", "Ciudad"]] = df["University name"].str.split(',',1, expand=True)
```

In [13]:
```
df.drop(columns = "University name")
```

Out[13]:

	Rank	locationLocation	Number of Studnet	Number of student per staffs	International Student	Female : male ratio	Universidad	Ciudad
0	1	United Kingdom	20.965	10.6	0.42	1900-01-02 00:52:00	University of Oxford	None
1	2	United States	21.887	9.6	0.25	1900-01-02 02:50:00	Harvard University	None
2	3	United Kingdom	20.185	11.3	0.39	1900-01-01 23:53:00	University of Cambridge	None
3	3	United States	16.164	7.1	0.24	1900-01-01 22:54:00	Stanford University	None
4	5	United States	11.415	8.2	0.33	40 : 60	Massachusetts Institute of Technology	None
...	...	...	...	...	...	...	...	...
2340	Reporter	United Kingdom	6.315	18.6	0.12	1900-01-02 17:35:00	York St John University	None
2341	Reporter	Nigeria	12.880	33.0	0	1900-01-02 00:52:00	Yusuf Maitama Sule University	Kano
2342	Reporter	Ukraine	3.869	15.4	0.01	34 : 66	Zhytomyr Polytechnic State University	None
2343	Reporter	Pakistan	4.906	8.8	0.01	1900-01-02 15:37:00	Ziauddin University	None
2344	Reporter	Jordan	5.768	18.1	0.32	1900-01-01 23:53:00	Zarqa University	None

2345 rows × 8 columns

los x index son las filas y con este codigo se eliminan las filas que no contengan datos (")

In [34]:
```
for x in df.index:
    if df.loc[x, "locationLocation"] == '':
        df.drop(x, inplace=True)
```

Este codigo rellena los espacios bacios o los espacios NaN

In [32]:
```
df=df.fillna('')
```

In [30]:
```
df["locationLocation"] = df["locationLocation"].str.replace('NaN','')
```

In [37]:
```
df
```

Out[37]:

	Rank	University name	locationLocation	Number of Studnet	Number of student per staffs	International Student	Female : male ratio	Universidad	Ciudad
0	1	University of Oxford	United Kingdom	20.965	10.6	0.42	1900-01-02 00:52:00	University of Oxford	
1	2	Harvard University	United States	21.887	9.6	0.25	1900-01-02 02:50:00	Harvard University	
2	3	University of Cambridge	United Kingdom	20.185	11.3	0.39	1900-01-01 23:53:00	University of Cambridge	
3	3	Stanford University	United States	16.164	7.1	0.24	1900-01-01 22:54:00	Stanford University	
4	5	Massachusetts Institute of Technology	United States	11.415	8.2	0.33	40 : 60	Massachusetts Institute of Technology	
...	...	...	...	...	...	...	...	...	...
2340	Reporter	York St John University	United Kingdom	6.315	18.6	0.12	1900-01-02 17:35:00	York St John University	
2341	Reporter	Yusuf Maitama Sule University, Kano	Nigeria	12.880	33.0	0	1900-01-02 00:52:00	Yusuf Maitama Sule University	Kano
2342	Reporter	Zhytomyr Polytechnic State University	Ukraine	3.869	15.4	0.01	34 : 66	Zhytomyr Polytechnic State University	
2343	Reporter	Ziauddin University	Pakistan	4.906	8.8	0.01	1900-01-02 15:37:00	Ziauddin University	
2344	Reporter	Zarqa University	Jordan	5.768	18.1	0.32	1900-01-01 23:53:00	Zarqa University	

2234 rows × 9 columns

Este codifgo resetea los index es decir los indices de las filas

In [42]:
```
df = df.reset_index(drop=True)
```

In [43]:
```
output_file = "datos_limpios_RankingUniversidadesDelMundo.xlsx"
df.to_excel(output_file, index=False)
```

In [ ]:
```

```