Pandas

Flexibility of Python Working with Big data

Importing pandas

In [11]:
```
import pandas as pd
```

Importing dataset

In [91]:
```
df = pd.read_csv('pokemon_data.csv')
print(df)
```

```
       #                    Name   Type 1  Type 2  HP  Attack  Defense  \
0      1               Bulbasaur    Grass  Poison  45      49       49
1      2                 Ivysaur    Grass  Poison  60      62       63
2      3                Venusaur    Grass  Poison  80      82       83
3      3  VenusaurMega Venusaur    Grass  Poison  80     100      123
4      4              Charmander     Fire     NaN  39      52       43
..   ...                     ...      ...     ...  ..     ...      ...
795  719                 Diancie     Rock   Fairy  50     100      150
796  719    DiancieMega Diancie     Rock   Fairy  50     160      110
797  720    HoopaHoopa Confined  Psychic   Ghost  80     110       60
798  720     HoopaHoopa Unbound  Psychic    Dark  80     160       60
799  721               Volcanion     Fire   Water  80     110      120

     Sp. Atk  Sp. Def  Speed  Generation  Legendary
0         65       65     45           1      False
1         80       80     60           1      False
2        100      100     80           1      False
3        122      120     80           1      False
4         60       50     65           1      False
..       ...      ...    ...         ...        ...
795      100      150     50           6       True
796      160      110    110           6       True
797      150      130     70           6       True
798      170      130     80           6       True
799      130       90     70           6       True

[800 rows x 12 columns]
```

In [13]:
```
df_excel = pd.read_excel('pokemon_data.xlsx')
print(df_excel)
```

```
       #                    Name   Type 1  Type 2  HP  Attack  Defense  \
0      1               Bulbasaur    Grass  Poison  45      49       49
1      2                 Ivysaur    Grass  Poison  60      62       63
2      3                Venusaur    Grass  Poison  80      82       83
3      3  VenusaurMega Venusaur    Grass  Poison  80     100      123
4      4              Charmander     Fire     NaN  39      52       43
..   ...                     ...      ...     ...  ..     ...      ...
795  719                Diancie     Rock   Fairy  50     100      150
796  719    DiancieMega Diancie     Rock   Fairy  50     160      110
797  720    HoopaHoopa Confined  Psychic   Ghost  80     110       60
798  720     HoopaHoopa Unbound  Psychic    Dark  80     160       60
799  721              Volcanion     Fire   Water  80     110      120

     Sp. Atk  Sp. Def  Speed  Generation  Legendary
0         65       65     45           1      False
1         80       80     60           1      False
2        100      100     80           1      False
3        122      120     80           1      False
4         60       50     65           1      False
..       ...      ...    ...         ...        ...
795      100      150     50           6       True
796      160      110    110           6       True
797      150      130     70           6       True
798      170      130     80           6       True
799      130       90     70           6       True

[800 rows x 12 columns]
```

In [14]:
```
df_text = pd.read_csv("pokemon_data.txt",delimiter='\t')
print(df_text)
```

```
       #                   Name   Type 1  Type 2  HP  Attack  Defense  \
0      1              Bulbasaur    Grass  Poison  45      49       49
1      2                Ivysaur    Grass  Poison  60      62       63
2      3               Venusaur    Grass  Poison  80      82       83
3      3  VenusaurMega Venusaur    Grass  Poison  80     100      123
4      4             Charmander     Fire     NaN  39      52       43
..   ...                    ...      ...     ...  ..     ...      ...
795  719                Diancie     Rock   Fairy  50     100      150
796  719    DiancieMega Diancie     Rock   Fairy  50     160      110
797  720    HoopaHoopa Confined  Psychic   Ghost  80     110       60
798  720     HoopaHoopa Unbound  Psychic    Dark  80     160       60
799  721              Volcanion     Fire   Water  80     110      120

     Sp. Atk  Sp. Def  Speed  Generation  Legendary
0         65       65     45           1      False
1         80       80     60           1      False
2        100      100     80           1      False
3        122      120     80           1      False
4         60       50     65           1      False
..       ...      ...    ...         ...        ...
795      100      150     50           6       True
796      160      110    110           6       True
797      150      130     70           6       True
798      170      130     80           6       True
799      130       90     70           6       True

[800 rows x 12 columns]
```

Reading Data in Pandas

Reading first and last rows of data

In [53]:
```
print(df.head(4))
print(df.tail(10))
```

```
   #                  Name Type 1  Type 2  HP  Attack  Defense  Sp. Atk  \
0  1             Bulbasaur  Grass  Poison  45      49       49       65
1  2               Ivysaur  Grass  Poison  60      62       63       80
2  3              Venusaur  Grass  Poison  80      82       83      100
3  3  VenusaurMega Venusaur  Grass  Poison  80     100      123      122

   Sp. Def  Speed  Generation  Legendary  Total
0       65     45           1      False    318
1       80     60           1      False    405
2      100     80           1      False    525
3      120     80           1      False    625
       #                 Name   Type 1  Type 2   HP  Attack  Defense  Sp. Atk  \
790  714               Noibat   Flying  Dragon   40      30       35       45
791  715              Noivern   Flying  Dragon   85      70       80       97
792  716              Xerneas    Fairy     NaN  126     131       95      131
793  717              Yveltal     Dark  Flying  126     131       95      131
794  718     Zygarde50% Forme   Dragon  Ground  108     100      121       81
795  719              Diancie     Rock   Fairy   50     100      150      100
796  719  DiancieMega Diancie     Rock   Fairy   50     160      110      160
797  720  HoopaHoopa Confined  Psychic   Ghost   80     110       60      150
798  720   HoopaHoopa Unbound  Psychic    Dark   80     160       60      170
799  721            Volcanion     Fire   Water   80     110      120      130

     Sp. Def  Speed  Generation  Legendary  Total
790       40     55           6      False    245
791       80    123           6      False    535
792       98     99           6       True    680
793       98     99           6       True    680
794       95     95           6       True    600
795      150     50           6       True    600
796      110    110           6       True    700
797      130     70           6       True    600
798      130     80           6       True    680
799       90     70           6       True    600
```

Reading columns

In [17]:
```
df.columns
```

Out[17]:
```
Index(['#', 'Name', 'Type 1', 'Type 2', 'HP', 'Attack', 'Defense', 'Sp. Atk',
       'Sp. Def', 'Speed', 'Generation', 'Legendary'],
      dtype='object')
```

Read each column

In [20]:
```
df[['Name','Type 1','Type 2']]
```

Out[20]:

	Name	Type 1	Type 2
0	Bulbasaur	Grass	Poison
1	Ivysaur	Grass	Poison
2	Venusaur	Grass	Poison
3	VenusaurMega Venusaur	Grass	Poison
4	Charmander	Fire	NaN
...	...	...	...
795	Diancie	Rock	Fairy
796	DiancieMega Diancie	Rock	Fairy
797	HoopaHoopa Confined	Psychic	Ghost
798	HoopaHoopa Unbound	Psychic	Dark
799	Volcanion	Fire	Water

800 rows × 3 columns

Reading each row

In [22]:
```
df.iloc[1:4]
```

Out[22]:

	#	Name	Type 1	Type 2	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Le
1	2	Ivysaur	Grass	Poison	60	62	63	80	80	60	1	
2	3	Venusaur	Grass	Poison	80	82	83	100	100	80	1	
3	3	VenusaurMega Venusaur	Grass	Poison	80	100	123	122	120	80	1	

Some part of data

In [23]:
```
df.iloc[1:4,5:22]  #[row,column] --> [row1:row2,column5:column21]
```

Out[23]:

	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legendary
1	62	63	80	80	60	1	False
2	82	83	100	100	80	1	False
3	100	123	122	120	80	1	False

Iterating in data

In [29]:
```
# for index, row in df.iterrows():
#     print(index,row['Name'])
```

Sorting and Describing data

In [44]:
```
df.describe()
```

Out[44]:

	#	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed
count	800.000000	800.000000	800.000000	800.000000	800.000000	800.000000	800.000000
mean	362.813750	69.258750	79.001250	73.842500	72.820000	71.902500	68.277500
std	208.343798	25.534669	32.457366	31.183501	32.722294	27.828916	29.060474
min	1.000000	1.000000	5.000000	5.000000	10.000000	20.000000	5.000000
25%	184.750000	50.000000	55.000000	50.000000	49.750000	50.000000	45.000000
50%	364.500000	65.000000	75.000000	70.000000	65.000000	70.000000	65.000000
75%	539.250000	80.000000	100.000000	90.000000	95.000000	90.000000	90.000000
max	721.000000	255.000000	190.000000	230.000000	194.000000	230.000000	180.000000

In [51]:
```
df.sort_values(['Type 1','HP'],ascending=[1,0])  #[1,0] means 1st column is ascedin
```

Out[51]:

	#	Name	Type 1	Type 2	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generat
520	469	Yanmega	Bug	Flying	86	76	86	116	56	95	
698	637	Volcarona	Bug	Fire	85	60	65	135	105	100	
231	214	Heracross	Bug	Fighting	80	125	75	40	95	85	
232	214	HeracrossMega Heracross	Bug	Fighting	80	185	115	40	105	75	
678	617	Accelgor	Bug	NaN	80	70	40	100	60	145	
...	...	...	...	...	...	...	...	...	...	...	
106	98	Krabby	Water	NaN	30	105	90	25	25	50	
125	116	Horsea	Water	NaN	30	40	70	70	25	60	
129	120	Staryu	Water	NaN	30	45	55	70	55	85	
139	129	Magikarp	Water	NaN	20	10	55	15	20	80	
381	349	Feebas	Water	NaN	20	15	20	10	55	80	

800 rows × 12 columns

Making changes to the data

Adding a column into the data

In [55]:
```
df['Total'] = df['HP']+df['Attack']+df['Defense']+df['Sp. Atk']+df['Sp. Def']+df['S
df.head(5)
```

Out[55]:

	#	Name	Type 1	Type 2	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Le
0	1	Bulbasaur	Grass	Poison	45	49	49	65	65	45	1	
1	2	Ivysaur	Grass	Poison	60	62	63	80	80	60	1	
2	3	Venusaur	Grass	Poison	80	82	83	100	100	80	1	
3	3	VenusaurMega Venusaur	Grass	Poison	80	100	123	122	120	80	1	
4	4	Charmander	Fire	NaN	39	52	43	60	50	65	1	

Droping the column

In [58]:
```
df = df.drop(columns = ['Total'])
df.head(2)
```

Out[58]:

	#	Name	Type 1	Type 2	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legend
0	1	Bulbasaur	Grass	Poison	45	49	49	65	65	45	1	F
1	2	Ivysaur	Grass	Poison	60	62	63	80	80	60	1	F

In [93]:
```
df['Total'] = df.iloc[:,4:10].sum(axis=1)  #axis =1 means it will add horizontally
df.head(3)
```

Out[93]:

	#	Name	Type 1	Type 2	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legend
0	1	Bulbasaur	Grass	Poison	45	49	49	65	65	45	1	F
1	2	Ivysaur	Grass	Poison	60	62	63	80	80	60	1	F
2	3	Venusaur	Grass	Poison	80	82	83	100	100	80	1	F

In [62]:
```
#We we have to move the added column to some other column place
cols = list(df.columns)
df = df[cols[0:4]+[cols[-1]]+cols[4:12]]
df.head(2)
```

Out[62]:

	#	Name	Type 1	Type 2	Total	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation
0	1	Bulbasaur	Grass	Poison	318	45	49	49	65	65	45	1
1	2	Ivysaur	Grass	Poison	405	60	62	63	80	80	60	1

Saving the data

In [64]:
```
df.to_csv('pokemon_data_total.csv',index = False)
```

In [67]:
```
df.to_excel('pokemon_data_md.xlsx',index=False)
```

In [68]:
```
df.to_csv('pokemon_data_md.txt',sep="\t")
```

Filtering Data

In [79]:
```
df_hp = df.loc[(df['HP']>120) & (df['Type 1']=='Water')] #Here and is % or is |
```

In [80]:
```
df_hp.head(3)
```

Out[80]:

	#	Name	Type 1	Type 2	Total	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Genera
142	131	Lapras	Water	Ice	535	130	85	80	85	95	60	
145	134	Vaporeon	Water	NaN	525	130	65	60	110	95	65	
185	171	Lanturn	Water	Electric	460	125	58	58	76	76	67	

In [81]:
```
#If the obtain data index is clumsy we can use reset_index
df_hp.reset_index(drop = True,inplace = True)
df_hp
```

Out[81]:

	#	Name	Type 1	Type 2	Total	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generat
0	131	Lapras	Water	Ice	535	130	85	80	85	95	60	
1	134	Vaporeon	Water	NaN	525	130	65	60	110	95	65	
2	171	Lanturn	Water	Electric	460	125	58	58	76	76	67	
3	320	Wailmer	Water	NaN	400	130	70	35	70	35	60	
4	321	Wailord	Water	NaN	500	170	90	45	90	45	60	
5	594	Alomomola	Water	NaN	470	165	75	80	40	45	65	

In [83]:
```
#Some examples
#If you want to get the data that contain particular string
df.loc[df['Name'].str.contains('Mega')].head(3)
```

Out[83]:

	#	Name	Type 1	Type 2	Total	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Genera
3	3	VenusaurMega Venusaur	Grass	Poison	625	80	100	123	122	120	80	
7	6	CharizardMega Charizard X	Fire	Dragon	634	78	130	111	130	85	100	
8	6	CharizardMega Charizard Y	Fire	Flying	634	78	104	78	159	115	100	

In [87]:
```
#Using regular expression
import re
df.loc[df['Type 1'].str.contains('n[a-z]*',flags = re.I,regex = True)] #flags can g
```

Out[87]:

	#	Name	Type 1	Type 2	Total	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	G
20	16	Pidgey	Normal	Flying	251	40	45	40	35	35	56	
21	17	Pidgeotto	Normal	Flying	349	63	60	55	50	50	71	
22	18	Pidgeot	Normal	Flying	479	83	80	75	70	70	101	
23	18	PidgeotMega Pidgeot	Normal	Flying	579	83	80	80	135	80	121	
24	19	Rattata	Normal	NaN	253	30	56	35	25	35	72	
...	...	...	...	...	...	...	...	...	...	...	...	...
775	705	Sliggoo	Dragon	NaN	452	68	75	53	83	113	60	
776	706	Goodra	Dragon	NaN	600	90	100	70	110	150	80	
790	714	Noibat	Flying	Dragon	245	40	30	35	45	40	55	
791	715	Noivern	Flying	Dragon	535	85	70	80	97	80	123	
794	718	Zygarde50% Forme	Dragon	Ground	600	108	100	121	81	95	95	

221 rows × 13 columns

Conditional Changes

In [94]:
```
df.loc[df['Total']> 500, ['Generation','Legendary']] = ['Test 1','Test 2']  #df.loc
```

```
C:\Users\saite\AppData\Local\Temp\ipykernel_26464\89345141.py:1: FutureWarning: Sett
ing an item of incompatible dtype is deprecated and will raise in a future error of
pandas. Value 'Test 1' has dtype incompatible with int64, please explicitly cast to
a compatible dtype first.
  df.loc[df['Total']> 500, ['Generation','Legendary']] = ['Test 1','Test 2']
C:\Users\saite\AppData\Local\Temp\ipykernel_26464\89345141.py:1: FutureWarning: Sett
ing an item of incompatible dtype is deprecated and will raise in a future error of
pandas. Value 'Test 2' has dtype incompatible with bool, please explicitly cast to a
compatible dtype first.
  df.loc[df['Total']> 500, ['Generation','Legendary']] = ['Test 1','Test 2']
```

In [95]:
```
df
```

Out[95]:

	#	Name	Type 1	Type 2	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generati
0	1	Bulbasaur	Grass	Poison	45	49	49	65	65	45	
1	2	Ivysaur	Grass	Poison	60	62	63	80	80	60	
2	3	Venusaur	Grass	Poison	80	82	83	100	100	80	Tes
3	3	VenusaurMega Venusaur	Grass	Poison	80	100	123	122	120	80	Tes
4	4	Charmander	Fire	NaN	39	52	43	60	50	65	
...	...	...	...	...	...	...	...	...	...	...	
795	719	Diancie	Rock	Fairy	50	100	150	100	150	50	Tes
796	719	DiancieMega Diancie	Rock	Fairy	50	160	110	160	110	110	Tes
797	720	HoopaHoopa Confined	Psychic	Ghost	80	110	60	150	130	70	Tes
798	720	HoopaHoopa Unbound	Psychic	Dark	80	160	60	170	130	80	Tes
799	721	Volcanion	Fire	Water	80	110	120	130	90	70	Tes

800 rows × 13 columns

Aggregate Statistics (Group by)

In [152…
```
df.groupby('Type 1')[['HP','Attack']].mean().sort_values('HP')  # groupby(columns_t
```

Out[152...

Type 1	HP	Attack
Bug	56.884058	70.971014
Electric	59.795455	69.090909
Ghost	64.437500	73.781250
Steel	65.222222	92.703704
Rock	65.363636	92.863636
Dark	66.806452	88.387097
Poison	67.250000	74.678571
Grass	67.271429	73.214286
Fighting	69.851852	96.777778
Fire	69.903846	84.769231
Psychic	70.631579	71.456140
Flying	70.750000	78.750000
Ice	72.000000	72.750000
Water	72.062500	74.151786
Ground	73.781250	95.750000
Fairy	74.117647	61.529412
Normal	77.275510	73.469388
Dragon	83.312500	112.125000

In [153...

```
df.groupby('Type 1')[['HP','Attack']].sum()
```

Out[153...

Type 1	HP	Attack
Bug	3925	4897
Dark	2071	2740
Dragon	2666	3588
Electric	2631	3040
Fairy	1260	1046
Fighting	1886	2613
Fire	3635	4408
Flying	283	315
Ghost	2062	2361
Grass	4709	5125
Ground	2361	3064
Ice	1728	1746
Normal	7573	7200
Poison	1883	2091
Psychic	4026	4073
Rock	2876	4086
Steel	1761	2503
Water	8071	8305

In [154...

```
df.groupby(['Type 1','Type 2'])[['HP','Attack']].count()
```

Out[154…

Type 1	Type 2	HP	Attack
Bug	Electric	2	2
	Fighting	2	2
	Fire	2	2
	Flying	14	14
	Ghost	1	1
...	...	...	...
Water	Ice	3	3
	Poison	3	3
	Psychic	5	5
	Rock	4	4
	Steel	1	1

136 rows × 2 columns

Working with large amount of data

In [160…

```
count = 0
for df in pd.read_csv('pokemon_data.csv',chunksize = 5):
    count += 1
print(count)
```

```
160
```

In [164…

```
new_df = pd.DataFrame(columns = df.columns)
for df in pd.read_csv('pokemon_data.csv',chunksize = 5):
    results = df.groupby('Type 1').count()
    new_df = pd.concat([new_df,results])
new_df.head(5)
new_df.shape
```

Out[164… (433, 12)

In [ ]: