Pandas

Import Pandas and Load Data Frame

- Read from csv file, and print first 5 rows.

```
import pandas as pd
#Read csv file
df = pd.read_csv('/content/pokemon_data.csv')
df.head(5)
```

	#	Name	Type 1	Type 2	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	L
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	

- Read from excel file, and print first 5 rows.

```
#Read excel file
df_xlsx = pd.read_excel('/content/pokemon_data.xlsx')
df_xlsx.head(5)
```

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

- Read from text file, and print first 5 rows.

```
#Read txt file
df=pd.read_csv('/content/pokemon_data.txt',delimiter='\t')
df.head(5)
```

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

Reading Data

- Reading Headers

```
#Reading headers
df.columns
```

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

- Read each column, by column name

```
#Read each column
df['Name']
```

```
0                    Bulbasaur
1                      Ivysaur
2                     Venusaur
3        VenusaurMega Venusaur
4                   Charmander
                  ...
795                    Diancie
796         DiancieMega Diancie
797        HoopaHoopa Confined
798         HoopaHoopa Unbound
799                   Volcanion
Name: Name, Length: 800, dtype: object
```

- Read specific element

```
#Read specific element
df['Name'][1]
```

```
    Ivysaur
```

- Read specific column,from l to r-1th row

```
df['Name'][0:5]
```

```
0                    Bulbasaur
1                      Ivysaur
2                     Venusaur
3        VenusaurMega Venusaur
4                   Charmander
Name: Name, dtype: object
```

- Read each row, by row number

```
#Read each row
df.iloc[2]
```

```
#                    3
Name           Venusaur
Type 1            Grass
Type 2           Poison
HP                   80
Attack               82
Defense              83
Sp. Atk             100
Sp. Def             100
Speed                80
Generation            1
Legendary         False
Name: 2, dtype: object
```

- Read rows from l to r-1

```
#Read rows from l to r-1
df.iloc[1:5]
```

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

- Read a specific location, using iloc

```
#Read specific location
df.iloc[2,1]
```

```
    'Venusaur'
```

- Iterate through each row

```
#Iterate through each row
for index, row in df.iterrows():
```

```
print(index, row['Name'])
if index==5:
 break
```

```
    0 Bulbasaur
    1 Ivysaur
    2 Venusaur
    3 VenusaurMega Venusaur
    4 Charmander
    5 Charmeleon
```

- Read all elements with the same condition. Use of loc

```
#Read all elements with same condition
df.loc[df['Type 1']=='Fire'].head(10)
```

	#	Name	Type 1	Type 2	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legendary
4	4	Charmander	Fire	NaN	39	52	43	60	50	65	1	False
5	5	Charmeleon	Fire	NaN	58	64	58	80	65	80	1	False
6	6	Charizard	Fire	Flying	78	84	78	109	85	100	1	False
7	6	CharizardMega Charizard X	Fire	Dragon	78	130	111	130	85	100	1	False
8	6	CharizardMega Charizard Y	Fire	Flying	78	104	78	159	115	100	1	False
42	37	Vulpix	Fire	NaN	38	41	40	50	65	65	1	False
43	38	Ninetales	Fire	NaN	73	76	75	81	100	100	1	False
63	58	Growlithe	Fire	NaN	55	70	45	70	50	60	1	False
64	59	Arcanine	Fire	NaN	90	110	80	100	80	95	1	False
83	77	Ponyta	Fire	NaN	50	85	55	65	65	90	1	False

▾ Describing and Sorting data

- Describing Data

```
#Describing Data
df.describe()
```

	#	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation
count	800.000000	800.000000	800.000000	800.000000	800.000000	800.000000	800.000000	800.00000
mean	362.813750	69.258750	79.001250	73.842500	72.820000	71.902500	68.277500	3.32375
std	208.343798	25.534669	32.457366	31.183501	32.722294	27.828916	29.060474	1.66129
min	1.000000	1.000000	5.000000	5.000000	10.000000	20.000000	5.000000	1.00000
25%	184.750000	50.000000	55.000000	50.000000	49.750000	50.000000	45.000000	2.00000
50%	364.500000	65.000000	75.000000	70.000000	65.000000	70.000000	65.000000	3.00000
75%	539.250000	80.000000	100.000000	90.000000	95.000000	90.000000	90.000000	5.00000
max	721.000000	255.000000	190.000000	230.000000	194.000000	230.000000	180.000000	6.00000

- Sort values, argument ascending=True/False

```
#Sort values
df.sort_values('Name', ascending=False)
```

#		Name	Type 1	Type 2	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legendary
794	718	Zygarde50% Forme	Dragon	Ground	108	100	121	81	95	95	6	True
695	634	Zweilous	Dark	Dragon	72	85	70	65	70	58	5	False
46	41	Zubat	Poison	Flying	40	45	35	30	40	55	1	False

- Sort with multiple parameters. First argument = a list of parameters, and the sec- ond argument list of boolean values if true or false

```
#Sort with multiple parameters
df.sort_values(['Type 1','HP'],
ascending=[False,True])
```

#		Name	Type 1	Type 2	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legendary
139	129	Magikarp	Water	NaN	20	10	55	15	20	80	1	False
381	349	Feebas	Water	NaN	20	15	20	10	55	80	3	False
97	90	Shellder	Water	NaN	30	65	100	45	25	40	1	False
106	98	Krabby	Water	NaN	30	105	90	25	25	50	1	False
125	116	Horsea	Water	NaN	30	40	70	70	25	60	1	False
...	...	...	...	...	...	...	...	...	...	...	...	...
232	214	HeracrossMega Heracross	Bug	Fighting	80	185	115	40	105	75	2	False
678	617	Accelgor	Bug	NaN	80	70	40	100	60	145	5	False
734	666	Vivillon	Bug	Flying	80	52	50	90	50	89	6	False
698	637	Volcarona	Bug	Fire	85	60	65	135	105	100	5	False
520	469	Yanmega	Bug	Flying	86	76	86	116	56	95	4	False

800 rows × 12 columns

▾ Making Changes to the data frame

- Adding new column to the data frame

```
#Adding a new column named 'Total'
df['Total'] = df['HP']+df['Attack']+df['Defense']+df['Speed']
df.head(5)
```

#		Name	Type 1	Type 2	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legendary	Total
0	1	Bulbasaur	Grass	Poison	45	49	49	65	65	45	1	False	188
1	2	Ivysaur	Grass	Poison	60	62	63	80	80	60	1	False	245
2	3	Venusaur	Grass	Poison	80	82	83	100	100	80	1	False	325
3	3	VenusaurMega Venusaur	Grass	Poison	80	100	123	122	120	80	1	False	383
4	4	Charmander	Fire	NaN	39	52	43	60	50	65	1	False	199

- Delete or drop specific column

```
#Drop/ Delete column = 'Total'
df = df.drop(columns=['Total'])
df.head(5)
```

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

- Add a column in a different way. By summing multiple columns to create a new one

```
#Another way to add column
df['Total']=df.iloc[:,4:10].sum(axis=1)
```

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

	#	Name	Type 1	Type 2	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legendary	Total
0	1	Bulbasaur	Grass	Poison	45	49	49	65	65	45	1	False	318
1	2	Ivysaur	Grass	Poison	60	62	63	80	80	60	1	False	405
2	3	Venusaur	Grass	Poison	80	82	83	100	100	80	1	False	525
3	3	VenusaurMega Venusaur	Grass	Poison	80	100	123	122	120	80	1	False	625
4	4	Charmander	Fire	NaN	39	52	43	60	50	65	1	False	309

- Rearranging columns

```
#Bringing last column to the 4th position
cols = list(df.columns)
df = df[cols[0:4]+[cols[-1]]+cols[4:12]]
df.head(5)
```

	#	Name	Type 1	Type 2	Total	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legendary
0	1	Bulbasaur	Grass	Poison	318	45	49	49	65	65	45	1	False
1	2	Ivysaur	Grass	Poison	405	60	62	63	80	80	60	1	False
2	3	Venusaur	Grass	Poison	525	80	82	83	100	100	80	1	False
3	3	VenusaurMega Venusaur	Grass	Poison	625	80	100	123	122	120	80	1	False
4	4	Charmander	Fire	NaN	309	39	52	43	60	50	65	1	False

Saving Data

- Save modifies df as modified.csv in desired location

```
#Saving data
df.to_csv('/content/modified.csv',index=False)
df.to_excel('/content/motified.xlsx', index=False)
#Separate all data with tabs
df.to_csv('/content/modified.csv',index=False, sep='\t')
```

Data filtering

- Filtering data based on multiple conditions

```
#Filtering Data based on multiple conditions
new_df = df.loc[(df['Type 1']=='Grass') & (df['Type 2']=='Poison') & (df['HP']>70)]
new_df
```

	#	Name	Type 1	Type 2	Total	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legendary
2	3	Venusaur	Grass	Poison	525	80	82	83	100	100	80	1	False
3	3	VenusaurMega Venusaur	Grass	Poison	625	80	100	123	122	120	80	1	False
50	45	Vileplume	Grass	Poison	490	75	80	85	110	90	50	1	False
77	71	Victreebel	Grass	Poison	490	80	105	65	100	70	70	1	False
652	591	Amoonguss	Grass	Poison	464	114	85	70	85	80	30	5	False

- Reset the index and drop the old index

```
#Reset Index
new_df = new_df.reset_index()
new_df
#To avoid old index, new_df.reset_index(drop = True)
```

	index	#		Name	Type 1	Type 2	Total	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legendary
0	2	3		Venusaur	Grass	Poison	525	80	82	83	100	100	80	1	False

- Regex Filtering (Filter based on textual patterns)

| 2 | 50 | 45 | | Vileplume | Grass | Poison | 490 | 75 | 80 | 85 | 110 | 90 | 50 | 1 | False |

```
#Pokemons whose Names have Mega as substring
df.loc[df['Name'].str.contains('Mega')].head(5)
```

	#	Name	Type 1	Type 2	Total	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legendary
3	3	VenusaurMega Venusaur	Grass	Poison	625	80	100	123	122	120	80	1	False
7	6	CharizardMega Charizard X	Fire	Dragon	634	78	130	111	130	85	100	1	False
8	6	CharizardMega Charizard Y	Fire	Flying	634	78	104	78	159	115	100	1	False
12	9	BlastoiseMega Blastoise	Water	NaN	630	79	103	120	135	115	78	1	False
19	15	BeedrillMega Beedrill	Bug	Poison	495	65	150	40	15	80	145	1	False

- For Filtering with multiple strings at once, we need to import regular expression

```
#Pkemons whose Type 1 is Fire or Grass
import re
df.loc[df['Type 1'].str.contains('Fire|Grass', regex = True)].head(10)
```

	#	Name	Type 1	Type 2	Total	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legendary
0	1	Bulbasaur	Grass	Poison	318	45	49	49	65	65	45	1	False
1	2	Ivysaur	Grass	Poison	405	60	62	63	80	80	60	1	False
2	3	Venusaur	Grass	Poison	525	80	82	83	100	100	80	1	False
3	3	VenusaurMega Venusaur	Grass	Poison	625	80	100	123	122	120	80	1	False
4	4	Charmander	Fire	NaN	309	39	52	43	60	50	65	1	False
5	5	Charmeleon	Fire	NaN	405	58	64	58	80	65	80	1	False
6	6	Charizard	Fire	Flying	534	78	84	78	109	85	100	1	False
7	6	CharizardMega Charizard X	Fire	Dragon	634	78	130	111	130	85	100	1	False
8	6	CharizardMega Charizard Y	Fire	Flying	634	78	104	78	159	115	100	1	False
42	37	Vulpix	Fire	NaN	299	38	41	40	50	65	65	1	False

- Find strings that start with specific group of letters.

```
#Pokemons whose name starts with "Pi"
df.loc[df['Name'].str.contains('^pi[a-z]*', flags=re.I, regex =True)].head(5)
#flags = re.I is for avoiding case.
```

	#	Name	Type 1	Type 2	Total	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legendary
20	16	Pidgey	Normal	Flying	251	40	45	40	35	35	56	1	False
21	17	Pidgeotto	Normal	Flying	349	63	60	55	50	50	71	1	False
22	18	Pidgeot	Normal	Flying	479	83	80	75	70	70	101	1	False
23	18	PidgeotMega Pidgeot	Normal	Flying	579	83	80	80	135	80	121	1	False
30	25	Pikachu	Electric	NaN	320	35	55	40	50	50	90	1	False

Conditional Changes

- Change one column of the data frame, if some condition is true.

```
#change all Type 1, that are Fire to Flamer
df.loc[df['Type 1']=='Fire', 'Type 1']='Flamer'
df.head(10)
```

	#	Name	Type 1	Type 2	Total	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legendary
0	1	Bulbasaur	Grass	Poison	318	45	49	49	65	65	45	1	False
1	2	Ivysaur	Grass	Poison	405	60	62	63	80	80	60	1	False
2	3	Venusaur	Grass	Poison	525	80	82	83	100	100	80	1	False
3	3	VenusaurMega Venusaur	Grass	Poison	625	80	100	123	122	120	80	1	False
4	4	Charmander	Flamer	NaN	309	39	52	43	60	50	65	1	False
5	5	Charmeleon	Flamer	NaN	405	58	64	58	80	65	80	1	False
6	6	Charizard	Flamer	Flying	534	78	84	78	109	85	100	1	False

- If Type 1 == 'Flamer' change 'Legendary' to True

| 8 | 6 | CharizardMega Charizard Y | Flamer | Flying | 634 | 78 | 104 | 78 | 159 | 115 | 100 | 1 | False |

```
df.loc[df['Type 1']=='Flamer', 'Legendery']=True
df.head(10)
```

	#	Name	Type 1	Type 2	Total	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legendary	Legendery
0	1	Bulbasaur	Grass	Poison	318	45	49	49	65	65	45	1	False	NaN
1	2	Ivysaur	Grass	Poison	405	60	62	63	80	80	60	1	False	NaN
2	3	Venusaur	Grass	Poison	525	80	82	83	100	100	80	1	False	NaN
3	3	VenusaurMega Venusaur	Grass	Poison	625	80	100	123	122	120	80	1	False	NaN
4	4	Charmander	Flamer	NaN	309	39	52	43	60	50	65	1	False	True
5	5	Charmeleon	Flamer	NaN	405	58	64	58	80	65	80	1	False	True
6	6	Charizard	Flamer	Flying	534	78	84	78	109	85	100	1	False	True
7	6	CharizardMega Charizard X	Flamer	Dragon	634	78	130	111	130	85	100	1	False	True
8	6	CharizardMega Charizard Y	Flamer	Flying	634	78	104	78	159	115	100	1	False	True
9	7	Squirtle	Water	NaN	314	44	48	65	50	64	43	1	False	NaN

- Change multiple columns for singel condition

```
#If Type 1= Flamer change Type 1 to Fire and Legentary to False
df.loc[df['Type 1']=='Flamer', ['Type 1','Legendary']]=['Fire','False']
df.head(10)
```

	#	Name	Type 1	Type 2	Total	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legendary	Legendery
0	1	Bulbasaur	Grass	Poison	318	45	49	49	65	65	45	1	False	NaN
1	2	Ivysaur	Grass	Poison	405	60	62	63	80	80	60	1	False	NaN
2	3	Venusaur	Grass	Poison	525	80	82	83	100	100	80	1	False	NaN
3	3	VenusaurMega Venusaur	Grass	Poison	625	80	100	123	122	120	80	1	False	NaN
4	4	Charmander	Fire	NaN	309	39	52	43	60	50	65	1	False	True
5	5	Charmeleon	Fire	NaN	405	58	64	58	80	65	80	1	False	True
6	6	Charizard	Fire	Flying	534	78	84	78	109	85	100	1	False	True
7	6	CharizardMega Charizard X	Fire	Dragon	634	78	130	111	130	85	100	1	False	True
8	6	CharizardMega Charizard Y	Fire	Flying	634	78	104	78	159	115	100	1	False	True
9	7	Squirtle	Water	NaN	314	44	48	65	50	64	43	1	False	NaN

Aggregate Statistics (Groupby)

- The functions are mean(), sum, count etc

```
#Count
df['count']=1
df.groupby(['Type 1']).count()['count']
```

```
Type 1
Bug        69
Dark       31
Dragon     32
Electric   44
Fairy      17
Fighting   27
```

```
Fire        52
Flying       4
Ghost       32
Grass       70
Ground      32
Ice         24
Normal      98
Poison      28
Psychic     57
Rock        44
Steel       27
Water      112
Name: count, dtype: int64
```

```
#Count of subsets of a set
df.groupby(['Type 1', 'Type 2']).count()['count']
```

```
Type 1  Type 2
Bug     Electric    2
        Fighting    2
        Fire        2
        Flying     14
        Ghost       1
                   ..
Water   Ice         3
        Poison      3
        Psychic     5
        Rock        4
        Steel       1
Name: count, Length: 136, dtype: int64
```

▾ Working with large amounts of data

- Seperate data into handaleable chunk sizes

```
#Setting chunksize
for df in pd.read_csv('/content/pokemon_data.csv', chunksize=5):
  print("Chunk df")
  print(df)
```

```
Chunk df
    #            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
4   4      Charmander     Fire     NaN  39      52       43       60

   Sp. Def  Speed  Generation  Legendary
0       65     45           1      False
1       80     60           1      False
2      100     80           1      False
3      120     80           1      False
4       50     65           1      False
Chunk df
    #            Name   Type 1  Type 2  HP  Attack  Defense  Sp. Atk  \
5   5      Charmeleon     Fire     NaN  58      64       58       80
6   6       Charizard     Fire  Flying  78      84       78      109
7   6  CharizardMega Charizard X  Fire  Dragon  78     130      111      130
8   6  CharizardMega Charizard Y  Fire  Flying  78     104       78      159
9   7        Squirtle    Water     NaN  44      48       65       50

   Sp. Def  Speed  Generation  Legendary
5       65     80           1      False
6       85    100           1      False
7       85    100           1      False
8      115    100           1      False
9       64     43           1      False
Chunk df
     #            Name  Type 1  Type 2  HP  Attack  Defense  Sp. Atk  \
10   8       Wartortle   Water     NaN  59      63       80       65
11   9       Blastoise   Water     NaN  79      83      100       85
12   9  BlastoiseMega Blastoise  Water     NaN  79     103      120      135
13  10        Caterpie     Bug     NaN  45      30       35       20
14  11         Metapod     Bug     NaN  50      20       55       25

    Sp. Def  Speed  Generation  Legendary
10       80     58           1      False
11      105     78           1      False
12      115     78           1      False
13       20     45           1      False
14       25     30           1      False
Chunk df
     #            Name  Type 1  Type 2  HP  Attack  Defense  Sp. Atk  \
15  12      Butterfree     Bug  Flying  60      45       50       90
```

```
16  13                  Weedle    Bug   Poison  40    35      30      20
17  14                  Kakuna    Bug   Poison  45    25      50      25
18  15                  Beedrill  Bug   Poison  65    90      40      45
19  15  BeedrillMega Beedrill     Bug   Poison  65    150     40      15

    Sp. Def  Speed  Generation  Legendary
15       80     70           1      False
16       20     50           1      False
17       25     35           1      False
18       80     75           1      False
19       80    145           1      False
Chunk df
     #                    Name  Type 1  Type 2  HP  Attack  Defense  Sp. Atk  \
```