Data Gethering

In csv format

1.Opening a local csv file

In [4]:

```
import pandas as pd
df = pd.read_csv("placement.csv")
```

In [5]:

```
df.head(1)
```

Out[5]:

	Unnamed: 0	cgpa	iq	placement
0	0	6.8	123.0	1

In [ ]:

```

```

2.Opening a csv file from an URL

in url i give github url (click on raw in github)

In [17]:

```
import requests
from io import StringIO

url = "https://raw.githubusercontent.com/campusx-official/100-days-of-machine-learning/ma
in/day15%20-%20working%20with%20csv%20files/test.csv"
headers = {"User-Agent": "Mozilla/5.0 (Macintosh; Intel Mac OS X 10.14; rv:66.0) Gecko/2
0100101 Firefox/66.0"}
req = requests.get(url, headers=headers)
data = StringIO(req.text)
#------------------------------------------------
df1 = pd.read_csv(data)
```

In [19]:

```
df1.head()
```

Out[19]:

	Unnamed: 0	Unnamed: 1	Unnamed: 2	Unnamed: 3	Unnamed: 4	Unnamed: 5	Unnamed: 6	Unname
0	0	enrollee_id	city	city_development_index	gender	relevent_experience	enrolled_university	education_le
1	1	29725	city_40	0.776	Male	No relevent experience	no_enrollment	Gradu
2	2	11561	city_21	0.624	NaN	No relevent experience	Full time course	Gradu
3	3	33241	city_115	0.789	NaN	No relevent experience	NaN	Gradu

Unnamed: 0	Unnamed: 666	Unnamed: city_162	Unnamed: 3	Unnamed: Mal	Has relevent experience	Unnamed: 6	Unname Mast
4	6	1	2	0.767	4	no_enrollment	

In [ ]:

3.Seperator

give own column name

In [22]:

```
import requests
from io import StringIO

url = "https://raw.githubusercontent.com/campusx-official/100-days-of-machine-learning/main/day15%20-%20working%20with%20csv%20files/movie_titles_metadata.tsv"
headers = {"User-Agent": "Mozilla/5.0 (Macintosh; Intel Mac OS X 10.14; rv:66.0) Gecko/20100101 Firefox/66.0"}
req = requests.get(url, headers=headers)
data = StringIO(req.text)
#-----------------------------------------------
movie = pd.read_csv(data,sep="\t",names=["sno","movie_name","release_year","rating","votes","geners"])
```

In [23]:

```
movie
```

Out[23]:

	sno	movie_name	release_year	rating	votes	geners
0	m0	10 things i hate about you	1999	6.9	62847.0	['comedy' 'romance']
1	m1	1492: conquest of paradise	1992	6.2	10421.0	['adventure' 'biography' 'drama' 'history']
2	m2	15 minutes	2001	6.1	25854.0	['action' 'crime' 'drama' 'thriller']
3	m3	2001: a space odyssey	1968	8.4	163227.0	['adventure' 'mystery' 'sci-fi']
4	m4	48 hrs.	1982	6.9	22289.0	['action' 'comedy' 'crime' 'drama' 'thriller']
...	...	...	...	...	...	...
612	m612	watchmen	2009	7.8	135229.0	['action' 'crime' 'fantasy' 'mystery' 'sci-fi'...
613	m613	xxx	2002	5.6	53505.0	['action' 'adventure' 'crime']
614	m614	x-men	2000	7.4	122149.0	['action' 'sci-fi']
615	m615	young frankenstein	1974	8.0	57618.0	['comedy' 'sci-fi']
616	m616	zulu dawn	1979	6.4	1911.0	['action' 'adventure' 'drama' 'history' 'war']

617 rows × 6 columns

In [25]:

```
movie.set_index("sno")
```

Out[25]:

sno	movie_name	release_year	rating	votes	geners
m0	10 things i hate about you	1999	6.9	62847.0	['comedy' 'romance']
m1	1492: conquest of paradise	1992	6.2	10421.0	['adventure' 'biography' 'drama' 'history']

	movie_name	release_year	rating	votes	genres
m2	15 minutes	2001	6.4	25654.0	['action' 'crime' 'drama' 'thriller']
m3	2001: a space odyssey	1968	8.4	163227.0	['adventure' 'mystery' 'sci-fi']
m4	48 hrs.	1982	6.9	22289.0	['action' 'comedy' 'crime' 'drama' 'thriller']
...	...	...	...	...	...
m612	watchmen	2009	7.8	135229.0	['action' 'crime' 'fantasy' 'mystery' 'sci-fi'...
m613	xxx	2002	5.6	53505.0	['action' 'adventure' 'crime']
m614	x-men	2000	7.4	122149.0	['action' 'sci-fi']
m615	young frankenstein	1974	8.0	57618.0	['comedy' 'sci-fi']
m616	zulu dawn	1979	6.4	1911.0	['action' 'adventure' 'drama' 'history' 'war']

617 rows × 5 columns

In [ ]:

4. Index_col Parameter

In [29]:

```
import requests
from io import StringIO

url = "https://raw.githubusercontent.com/campusx-official/100-days-of-machine-learning/main/day15%20-%20working%20with%20csv%20files/aug_train.csv"
headers = {"User-Agent": "Mozilla/5.0 (Macintosh; Intel Mac OS X 10.14; rv:66.0) Gecko/20100101 Firefox/66.0"}
req = requests.get(url, headers=headers)
data = StringIO(req.text)
#-----------------------------------------------
df2 = pd.read_csv(data,index_col="enrollee_id")
```

In [31]:

```
df2.head()
```

Out[31]:

enrollee_id	city	city_development_index	gender	relevent_experience	enrolled_university	education_level	major_discipline
8949	city_103	0.920	Male	Has relevent experience	no_enrollment	Graduate	STEM
29725	city_40	0.776	Male	No relevent experience	no_enrollment	Graduate	STEM
11561	city_21	0.624	NaN	No relevent experience	Full time course	Graduate	STEM
33241	city_115	0.789	NaN	No relevent experience	NaN	Graduate	Business Degree
666	city_162	0.767	Male	Has relevent experience	no_enrollment	Masters	STEM

In [ ]:

Header Parameter

when column is become one row or records

In [50]:

```
import requests
from io import StringIO

url = "https://raw.githubusercontent.com/campusx-official/100-days-of-machine-learning/main/day15%20-%20working%20with%20csv%20files/test.csv"
headers = {"User-Agent": "Mozilla/5.0 (Macintosh; Intel Mac OS X 10.14; rv:66.0) Gecko/20100101 Firefox/66.0"}
req = requests.get(url, headers=headers)
data = StringIO(req.text)
#-------------------------------------------
df3 = pd.read_csv(data,header=1)
```

In [75]:

```
df3
```

Out[75]:

	0	enrollee_id	city	city_development_index	gender	relevent_experience	enrolled_university	education_level	major_dis
0	1	29725	city_40	0.776	Male	No relevent experience	no_enrollment	Graduate	
1	2	11561	city_21	0.624	NaN	No relevent experience	Full time course	Graduate	
2	3	33241	city_115	0.789	NaN	No relevent experience	NaN	Graduate	Bu
3	4	666	city_162	0.767	Male	Has relevent experience	no_enrollment	Masters	

In [ ]:

```

```

5. use_cols parameter

Select whatever column you not want

In [70]:

```
import requests
from io import StringIO

url = "https://raw.githubusercontent.com/campusx-official/100-days-of-machine-learning/main/day15%20-%20working%20with%20csv%20files/aug_train.csv"
headers = {"User-Agent": "Mozilla/5.0 (Macintosh; Intel Mac OS X 10.14; rv:66.0) Gecko/20100101 Firefox/66.0"}
req = requests.get(url, headers=headers)
data = StringIO(req.text)
#-------------------------------------------------
df5 = pd.read_csv(data,usecols=["enrollee_id","city","relevent_experience"])
```

In [78]:

```
df5
```

Out[78]:

	enrollee_id	city	relevent_experience
0	8949	city_103	Has relevent experience

1	enrollee_id	city	No relevent experience
2	11561	city_21	No relevent experience
3	33241	city_115	No relevent experience
4	666	city_162	Has relevent experience
...	...	...	...
19153	7386	city_173	No relevent experience
19154	31398	city_103	Has relevent experience
19155	24576	city_103	Has relevent experience
19156	5756	city_65	Has relevent experience
19157	23834	city_67	No relevent experience

19158 rows × 3 columns

In [ ]:

Skip Rows

In [81]:

```
import requests
from io import StringIO

url = "https://raw.githubusercontent.com/campusx-official/100-days-of-machine-learning/main/day15%20-%20working%20with%20csv%20files/aug_train.csv"
headers = {"User-Agent": "Mozilla/5.0 (Macintosh; Intel Mac OS X 10.14; rv:66.0) Gecko/20100101 Firefox/66.0"}
req = requests.get(url, headers=headers)
data = StringIO(req.text)
#-----------------------------------------
df6 = pd.read_csv(data,skiprows=[0,1])
```

In [82]:

```
df6
```

Out[82]:

	29725	city_40	0.7759999999999999	Male	No relevent experience	no_enrollment	Graduate	STEM	15	50-99	Pvt Ltd	>
0	11561	city_21	0.624	NaN	No relevent experience	Full time course	Graduate	STEM	5	NaN	NaN	nev
1	33241	city_115	0.789	NaN	No relevent experience	NaN	Graduate	Business Degree	<1	NaN	Pvt Ltd	nev
2	666	city_162	0.767	Male	Has relevent experience	no_enrollment	Masters	STEM	>20	50-99	Funded Startup	
3	21651	city_176	0.764	NaN	Has relevent experience	Part time course	Graduate	STEM	11	NaN	NaN	
4	28806	city_160	0.920	Male	Has relevent experience	no_enrollment	High School	NaN	5	50-99	Funded Startup	

...	...	...	...	...	...	...	...	...	...	...	...	...
19151	29725 7386	city_40 city_173	0.7759999999999999 0.878	Male Male	No relevent experience Has relevent experience	no_enrollment no_enrollment	Graduate Graduate	STEM Humanities	15 14	50-99 NaN	Pvt Ltd NaN	>
19152	31398	city_103	0.920	Male	Has relevent experience	no_enrollment	Graduate	STEM	14	NaN	NaN	
19153	24576	city_103	0.920	Male	Has relevent experience	no_enrollment	Graduate	STEM	>20	50-99	Pvt Ltd	
19154	5756	city_65	0.802	Male	Has relevent experience	no_enrollment	High School	NaN	<1	500-999	Pvt Ltd	
19155	23834	city_67	0.855	NaN	No relevent experience	no_enrollment	Primary School	NaN	2	NaN	NaN	

19156 rows × 14 columns

In [ ]:

```
```

nrows

when you have lots of data and you select some pic of data

In [83]:

```
import requests
from io import StringIO

url = "https://raw.githubusercontent.com/campusx-official/100-days-of-machine-learning/main/day15%20-%20working%20with%20csv%20files/aug_train.csv"
headers = {"User-Agent": "Mozilla/5.0 (Macintosh; Intel Mac OS X 10.14; rv:66.0) Gecko/20100101 Firefox/66.0"}
req = requests.get(url, headers=headers)
data = StringIO(req.text)
#----------------------------------------------------
df7 = pd.read_csv(data,nrows=1000)
```

In [85]:

```
df7
```

Out[85]:

	enrollee_id	city	city_development_index	gender	relevent_experience	enrolled_university	education_level	major_dis
0	8949	city_103	0.920	Male	Has relevent experience	no_enrollment	Graduate	
1	29725	city_40	0.776	Male	No relevent experience	no_enrollment	Graduate	
2	11561	city_21	0.624	NaN	No relevent experience	Full time course	Graduate	
3	33241	city_115	0.789	NaN	No relevent experience	NaN	Graduate	Bu
4	666	city_162	0.767	Male	Has relevent experience	no_enrollment	Masters	
...	...	...	...	...	...	...	...	
995	28500	city_21	0.624	NaN	Has relevent experience	Full time course	Graduate	

	enrollee_id	city	city_development_index	gender	relevent_experience	enrolled_university	education_level	major_dis
996	10371	city_103	0.920	Female	No relevent experience	no_enrollment	Phd	
997	10028	city_73	0.754	Male	Has relevent experience	no_enrollment	Graduate	
998	29671	city_40	0.776	Male	No relevent experience	Full time course	Graduate	
999	4482	city_103	0.920	Male	Has relevent experience	no_enrollment	Graduate	

1000 rows × 14 columns

In [ ]:

```
```

Encoding Parameter

Basically a lot of scanario you find UTF-8 encoding dataset but in some of the cases you find some different encoding

In [ ]:

```
pd.read_csv('zomato.csv',encoding='latin-1')
```

In [ ]:

```
```

Skip bad line

when you load dataset that time in your row find some extra thing some it give some error while loading to skip this error or load the dataset

In [ ]:

```
pd.read_csv('BX-Books.csv', sep=';', encoding="latin-1",error_bad_lines=False)
```

In [ ]:

```
```

Dtype Parameter

when your output variable is float like 0.0,1.0 and u want to convert this float values into int

In [95]:

```
import requests
from io import StringIO

url = "https://raw.githubusercontent.com/campusx-official/100-days-of-machine-learning/main/day15%20-%20working%20with%20csv%20files/aug_train.csv"
headers = {"User-Agent": "Mozilla/5.0 (Macintosh; Intel Mac OS X 10.14; rv:66.0) Gecko/20100101 Firefox/66.0"}
req = requests.get(url, headers=headers)
data = StringIO(req.text)
#-----------------------------------------------
```

```
df8 = pd.read_csv(data,dtype={"target":int})
```

In [96]:

```
# Now target column is int
df8.head(1)
```

Out[96]:

	enrollee_id	city	city_development_index	gender	relevent_experience	enrolled_university	education_level	major_discip
0	8949	city_103	0.92	Male	Has relevent experience	no_enrollment	Graduate	S

In [ ]:

```

```

Handling Dates

when you have dates column in you dataset and you want date column treate as a date not as string so we pass parse_dates parameter

In [ ]:

```
pd.read_csv('IPL Matches 2008-2020.csv',parse_dates=['date']).info()
```

Output

<class 'pandas.core.frame.DataFrame'> RangeIndex: 816 entries, 0 to 815 Data columns (total 17 columns):] Column Non-Null Count Dtype

---

0 id 816 non-null int64
1 city 803 non-null object
2 date 816 non-null datetime64[ns]

In [ ]:

```

```

na_values parameter

some of the dataset NAN value will dash(-) or double dash(--) so we say the our dataset to consoderd the NAN values to - and --

In [ ]:

```
pd.read_csv('aug_train.csv',,na_values=["-","--"])
```

In [ ]:

```

```

Loading Huge Data In Chunks

chunksize

In [102]:

```
df9 =pd.read_csv("placement.csv",chunksize=30)
```

In [103]:

```
for chunk in df9:
    print(chunk.shape)
```

```
(30, 4)
(30, 4)
(30, 4)
(10, 4)
```

In [111]:

```
for chucks in df9:
    chucks
```