Working on a CSV file

In [1]:

```
import pandas as pd
```

1. Opening a local CSV file

In [2]:

```
df = pd.read_csv("aug_train.csv")
df
```

Out[2]:

	enrollee_id	city	city_development_index	gender	relevent_experience	enrolled_university	education_level	major_discipline
0	8949	city_103	0.920	Male	Has relevent experience	no_enrollment	Graduate	STEM
1	29725	city_40	0.776	Male	No relevent experience	no_enrollment	Graduate	STEM
2	11561	city_21	0.624	NaN	No relevent experience	Full time course	Graduate	STEM
3	33241	city_115	0.789	NaN	No relevent experience	NaN	Graduate	Business Degree
4	666	city_162	0.767	Male	Has relevent experience	no_enrollment	Masters	STEM
...	...	...	...	...	...	...	...	...
19153	7386	city_173	0.878	Male	No relevent experience	no_enrollment	Graduate	Humanities
19154	31398	city_103	0.920	Male	Has relevent experience	no_enrollment	Graduate	STEM
19155	24576	city_103	0.920	Male	Has relevent experience	no_enrollment	Graduate	STEM
19156	5756	city_65	0.802	Male	Has relevent experience	no_enrollment	High School	NaN
19157	23834	city_67	0.855	NaN	No relevent experience	no_enrollment	Primary School	NaN

19158 rows × 14 columns

2. Opening a CSV file from an URL

In [3]:

```
import requests
from io import StringIO

url = "https://raw.githubusercontent.com/cs109/2014_data/master/countries.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)

pd.read_csv(data)
```

Out[3]:

	Country	Region
0	Algeria	AFRICA
1	Angola	AFRICA
2	Benin	AFRICA
3	Botswana	AFRICA
4	Burkina	AFRICA
...	...	...
189	Paraguay	SOUTH AMERICA
190	Peru	SOUTH AMERICA
191	Suriname	SOUTH AMERICA
192	Uruguay	SOUTH AMERICA
193	Venezuela	SOUTH AMERICA

194 rows × 2 columns

3. Seperate Parameter

In [4]:

```
pd.read_csv("movie_titles_metadata.tsv")
```

Out[4]:

	m0\t10 things i hate about you\t1999\t6.90\t62847\t['comedy' 'romance']
0	m1\t1492: conquest of paradise\t1992\t6.20\t10...
1	m2\t15 minutes\t2001\t6.10\t25854\t['action' '...
2	m3\t2001: a space odyssey\t1968\t8.40\t163227\...
3	m4\t48 hrs.\t1982\t6.90\t22289\t['action' 'com...
4	m5\tthe fifth element\t1997\t7.50\t133756\t['a...
...	...
611	m612\twatchmen\t2009\t7.80\t135229\t['action' ...
612	m613\txxx\t2002\t5.60\t53505\t['action' 'adven...
613	m614\tx-men\t2000\t7.40\t122149\t['action' 'sc...
614	m615\tyoung frankenstein\t1974\t8.00\t57618\t[...
615	m616\tzulu dawn\t1979\t6.40\t1911\t['action' '...

616 rows × 1 columns

In [5]:

```
d.read_csv("movie_titles_metadata.tsv", sep = '\t', names =['S_No.', 'Name', 'Release_Year', 'Rating', 'Votes', 'Genres'])
```

Out[5]:

	S_No.	Name	Release_Year	Rating	Votes	Genres
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

4. Index Col Parameter

In [8]:

```
pd.read_csv("aug_train.csv")
```

Out[8]:

	enrollee_id	city	city_development_index	gender	relevent_experience	enrolled_university	education_level	major_discipline
0	8949	city_103	0.920	Male	Has relevent experience	no_enrollment	Graduate	STEM
1	29725	city_40	0.776	Male	No relevent experience	no_enrollment	Graduate	STEM
2	11561	city_21	0.624	NaN	No relevent experience	Full time course	Graduate	STEM
3	33241	city_115	0.789	NaN	No relevent experience	NaN	Graduate	Business Degree
4	666	city_162	0.767	Male	Has relevent experience	no_enrollment	Masters	STEM
...	...	...	...	...	...	...	...	...
19153	7386	city_173	0.878	Male	No relevent experience	no_enrollment	Graduate	Humanities
19154	31398	city_103	0.920	Male	Has relevent experience	no_enrollment	Graduate	STEM
19155	24576	city_103	0.920	Male	Has relevent experience	no_enrollment	Graduate	STEM
19156	5756	city_65	0.802	Male	Has relevent experience	no_enrollment	High School	NaN
19157	23834	city_67	0.855	NaN	No relevent experience	no_enrollment	Primary School	NaN

19158 rows × 14 columns

In [7]:

```
pd.read_csv("aug_train.csv", index_col = 'enrollee_id')
```

Out[7]:

enrollee_id	city	city_development_index	gender	relevent_experience	enrolled_university	education_level	major_discipline	experie
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	
...	...	...	...	...	...	...	...	
7386	city_173	0.878	Male	No relevent experience	no_enrollment	Graduate	Humanities	
31398	city_103	0.920	Male	Has relevent experience	no_enrollment	Graduate	STEM	
24576	city_103	0.920	Male	Has relevent experience	no_enrollment	Graduate	STEM	
5756	city_65	0.802	Male	Has relevent experience	no_enrollment	High School	NaN	
23834	city_67	0.855	NaN	No relevent experience	no_enrollment	Primary School	NaN	

19158 rows × 13 columns

5. Header Parameter

In [10]:

```
pd.read_csv("test.csv")
```

Out[10]:

	Unnamed: 0	Unnamed: 1	Unnamed: 2	Unnamed: 3	Unnamed: 4	Unnamed: 5	Unnamed: 6	Unnamed: 7	Unname
0	0	enrollee_id	city	city_development_index	gender	relevent_experience	enrolled_university	education_level	major_discip
1	1	29725	city_40	0.776	Male	No relevent experience	no_enrollment	Graduate	ST
2	2	11561	city_21	0.624	NaN	No relevent experience	Full time course	Graduate	ST
3	3	33241	city_115	0.789	NaN	No relevent experience	NaN	Graduate	Busir Deg
4	4	666	city_162	0.767	Male	Has relevent experience	no_enrollment	Masters	ST

In [12]:

```
pd.read_csv("test.csv", header = 1)
```

Out[12]:

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

6. use_cols parameter

In [13]:

```
pd.read_csv("aug_train.csv")
```

Out[13]:

	enrollee_id	city	city_development_index	gender	relevent_experience	enrolled_university	education_level	major_discipline
0	8949	city_103	0.920	Male	Has relevent experience	no_enrollment	Graduate	STEM
1	29725	city_40	0.776	Male	No relevent experience	no_enrollment	Graduate	STEM
2	11561	city_21	0.624	NaN	No relevent experience	Full time course	Graduate	STEM
3	33241	city_115	0.789	NaN	No relevent experience	NaN	Graduate	Business Degree
4	666	city_162	0.767	Male	Has relevent experience	no_enrollment	Masters	STEM
...	...	...	...	...	...	...	...	...
19153	7386	city_173	0.878	Male	No relevent experience	no_enrollment	Graduate	Humanities
19154	31398	city_103	0.920	Male	Has relevent experience	no_enrollment	Graduate	STEM
19155	24576	city_103	0.920	Male	Has relevent experience	no_enrollment	Graduate	STEM
19156	5756	city_65	0.802	Male	Has relevent experience	no_enrollment	High School	NaN
19157	23834	city_67	0.855	NaN	No relevent experience	no_enrollment	Primary School	NaN

19158 rows × 14 columns

In [14]:

```
pd.read_csv("aug_train.csv", usecols = ['enrollee_id', 'gender', 'education_level'])
```

Out[14]:

	enrollee_id	gender	education_level
0	8949	Male	Graduate
1	29725	Male	Graduate
2	11561	NaN	Graduate
3	33241	NaN	Graduate
4	666	Male	Masters
...	...	...	...
19153	7386	Male	Graduate
19154	31398	Male	Graduate
19155	24576	Male	Graduate
19156	5756	Male	High School
19157	23834	NaN	Primary School

19158 rows × 3 columns

7. Squeeze Parameter

In [15]:

```
pd.read_csv("aug_train.csv")
```

Out[15]:

	enrollee_id	city	city_development_index	gender	relevent_experience	enrolled_university	education_level	major_discipline
0	8949	city_103	0.920	Male	Has relevent experience	no_enrollment	Graduate	STEM
1	29725	city_40	0.776	Male	No relevent experience	no_enrollment	Graduate	STEM
2	11561	city_21	0.624	NaN	No relevent experience	Full time course	Graduate	STEM
3	33241	city_115	0.789	NaN	No relevent experience	NaN	Graduate	Business Degree
4	666	city_162	0.767	Male	Has relevent experience	no_enrollment	Masters	STEM
...	...	...	...	...	...	...	...	...
19153	7386	city_173	0.878	Male	No relevent experience	no_enrollment	Graduate	Humanities
19154	31398	city_103	0.920	Male	Has relevent experience	no_enrollment	Graduate	STEM
19155	24576	city_103	0.920	Male	Has relevent experience	no_enrollment	Graduate	STEM
19156	5756	city_65	0.802	Male	Has relevent experience	no_enrollment	High School	NaN
19157	23834	city_67	0.855	NaN	No relevent experience	no_enrollment	Primary School	NaN

19158 rows × 14 columns

In [27]:

```
pd.read_csv("aug_train.csv",  usecols = ['gender'], squeeze = True)
```

```
C:\Users\TOSHIBA\AppData\Local\Temp\ipykernel_1544\295609253.py:1: FutureWarning: The squeeze argument has b
een deprecated and will be removed in a future version. Append .squeeze("columns") to the call to squeeze.


  pd.read_csv("aug_train.csv",  usecols = ['gender'], squeeze = True)
```

Out[27]:

```
0        Male
1        Male
2         NaN
3         NaN
4        Male
         ...
19153    Male
19154    Male
19155    Male
19156    Male
19157     NaN
Name: gender, Length: 19158, dtype: object
```

8. Skiprows/nrows Parameter

In [18]:

```
pd.read_csv("aug_train.csv")
```

Out[18]:

	enrollee_id	city	city_development_index	gender	relevent_experience	enrolled_university	education_level	major_discipline
0	8949	city_103	0.920	Male	Has relevent experience	no_enrollment	Graduate	STEM
1	29725	city_40	0.776	Male	No relevent experience	no_enrollment	Graduate	STEM
2	11561	city_21	0.624	NaN	No relevent experience	Full time course	Graduate	STEM
3	33241	city_115	0.789	NaN	No relevent experience	NaN	Graduate	Business Degree
4	666	city_162	0.767	Male	Has relevent experience	no_enrollment	Masters	STEM
...	...	...	...	...	...	...	...	...
19153	7386	city_173	0.878	Male	No relevent experience	no_enrollment	Graduate	Humanities
19154	31398	city_103	0.920	Male	Has relevent experience	no_enrollment	Graduate	STEM
19155	24576	city_103	0.920	Male	Has relevent experience	no_enrollment	Graduate	STEM
19156	5756	city_65	0.802	Male	Has relevent experience	no_enrollment	High School	NaN
19157	23834	city_67	0.855	NaN	No relevent experience	no_enrollment	Primary School	NaN

19158 rows × 14 columns

In [21]:

```
# Read CSV file and skip the first row
pd.read_csv('aug_train.csv', skiprows=1)
```

Out[21]:

	8949	city_103	0.92	Male	Has relevent experience	no_enrollment	Graduate	STEM	>20	Unnamed: 9	Unnamed: 10	1	36	1.0
0	29725	city_40	0.776	Male	No relevent experience	no_enrollment	Graduate	STEM	15	50-99	Pvt Ltd	>4	47	0.0
1	11561	city_21	0.624	NaN	No relevent experience	Full time course	Graduate	STEM	5	NaN	NaN	never	83	0.0
2	33241	city_115	0.789	NaN	No relevent experience	NaN	Graduate	Business Degree	<1	NaN	Pvt Ltd	never	52	1.0
3	666	city_162	0.767	Male	Has relevent experience	no_enrollment	Masters	STEM	>20	50-99	Funded Startup	4	8	0.0
4	21651	city_176	0.764	NaN	Has relevent experience	Part time course	Graduate	STEM	11	NaN	NaN	1	24	1.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
19152	7386	city_173	0.878	Male	No relevent experience	no_enrollment	Graduate	Humanities	14	NaN	NaN	1	42	1.0
19153	31398	city_103	0.920	Male	Has relevent experience	no_enrollment	Graduate	STEM	14	NaN	NaN	4	52	1.0
19154	24576	city_103	0.920	Male	Has relevent experience	no_enrollment	Graduate	STEM	>20	50-99	Pvt Ltd	4	44	0.0
19155	5756	city_65	0.802	Male	Has relevent experience	no_enrollment	High School	NaN	<1	500-999	Pvt Ltd	2	97	0.0
19156	23834	city_67	0.855	NaN	No relevent experience	no_enrollment	Primary School	NaN	2	NaN	NaN	1	127	0.0

19157 rows × 14 columns

In [23]:

```
# Read CSV file and skip the first two rows and the fifth row
pd.read_csv('aug_train.csv', skiprows=[0, 1, 4])
```

Out[23]:

	29725	city_40	0.7759999999999999	Male	No relevent experience	no_enrollment	Graduate	STEM	15	50-99	Pvt Ltd	>4	47	0.0
0	11561	city_21	0.624	NaN	No relevent experience	Full time course	Graduate	STEM	5	NaN	NaN	never	83	0.0
1	666	city_162	0.767	Male	Has relevent experience	no_enrollment	Masters	STEM	>20	50-99	Funded Startup	4	8	0.0
2	21651	city_176	0.764	NaN	Has relevent experience	Part time course	Graduate	STEM	11	NaN	NaN	1	24	1.0
3	28806	city_160	0.920	Male	Has relevent experience	no_enrollment	High School	NaN	5	50-99	Funded Startup	1	24	0.0
4	402	city_46	0.762	Male	Has relevent experience	no_enrollment	Graduate	STEM	13	<10	Pvt Ltd	>4	18	1.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
19150	7386	city_173	0.878	Male	No relevent experience	no_enrollment	Graduate	Humanities	14	NaN	NaN	1	42	1.0
19151	31398	city_103	0.920	Male	Has relevent experience	no_enrollment	Graduate	STEM	14	NaN	NaN	4	52	1.0
19152	24576	city_103	0.920	Male	Has relevent experience	no_enrollment	Graduate	STEM	>20	50-99	Pvt Ltd	4	44	0.0
19153	5756	city_65	0.802	Male	Has relevent experience	no_enrollment	High School	NaN	<1	500-999	Pvt Ltd	2	97	0.0
19154	23834	city_67	0.855	NaN	No relevent experience	no_enrollment	Primary School	NaN	2	NaN	NaN	1	127	0.0

19155 rows × 14 columns

In [25]:

```
# Read CSV file and skip rows from 2 to 4 (inclusive)
pd.read_csv('aug_train.csv', skiprows=range(2, 5))
```

Out[25]:

	enrollee_id	city	city_development_index	gender	relevent_experience	enrolled_university	education_level	major_discipline
0	8949	city_103	0.920	Male	Has relevent experience	no_enrollment	Graduate	STEM
1	666	city_162	0.767	Male	Has relevent experience	no_enrollment	Masters	STEM
2	21651	city_176	0.764	NaN	Has relevent experience	Part time course	Graduate	STEM
3	28806	city_160	0.920	Male	Has relevent experience	no_enrollment	High School	NaN
4	402	city_46	0.762	Male	Has relevent experience	no_enrollment	Graduate	STEM
...	...	...	...	...	...	...	...	...
19150	7386	city_173	0.878	Male	No relevent experience	no_enrollment	Graduate	Humanities
19151	31398	city_103	0.920	Male	Has relevent experience	no_enrollment	Graduate	STEM
19152	24576	city_103	0.920	Male	Has relevent experience	no_enrollment	Graduate	STEM
19153	5756	city_65	0.802	Male	Has relevent experience	no_enrollment	High School	NaN
19154	23834	city_67	0.855	NaN	No relevent experience	no_enrollment	Primary School	NaN

19155 rows × 14 columns

In [29]:

```
# Define a custom function to skip rows
def skip_multiple_of_5(index):
    return (index + 1) % 5 != 0

# Read the CSV file and apply the custom function to skip rows
pd.read_csv('aug_train.csv', skiprows=lambda x: skip_multiple_of_5(x))
```

Out[29]:

	33241	city_115	0.789	Unnamed: 3	No relevent experience	Unnamed: 5	Graduate	Business Degree	<1	Unnamed: 9	Pvt Ltd	never	52	1.0
0	27107	city_103	0.920	Male	Has relevent experience	no_enrollment	Graduate	STEM	7	50-99	Pvt Ltd	1	46	1.0
1	5826	city_21	0.624	Male	No relevent experience	NaN	NaN	NaN	2	NaN	NaN	never	24	0.0
2	2156	city_21	0.624	NaN	Has relevent experience	no_enrollment	Graduate	STEM	7	10000+	Pvt Ltd	never	23	1.0
3	7041	city_40	0.776	Male	Has relevent experience	no_enrollment	Graduate	Humanities	<1	1000-4999	Pvt Ltd	1	65	0.0
4	21538	city_100	0.887	Male	Has relevent experience	no_enrollment	High School	NaN	11	<10	Pvt Ltd	1	8	1.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
3825	25191	city_10	0.895	Other	Has relevent experience	no_enrollment	Graduate	STEM	16	10/49	Pvt Ltd	1	36	0.0
3826	19765	city_30	0.698	NaN	Has relevent experience	no_enrollment	Masters	STEM	11	100-500	Pvt Ltd	1	17	0.0
3827	33047	city_103	0.920	Male	Has relevent experience	no_enrollment	Graduate	STEM	>20	10000+	Pvt Ltd	>4	18	0.0
3828	9212	city_21	0.624	NaN	Has relevent experience	no_enrollment	Masters	STEM	3	100-500	Pvt Ltd	3	40	1.0
3829	7386	city_173	0.878	Male	No relevent experience	no_enrollment	Graduate	Humanities	14	NaN	NaN	1	42	1.0

3830 rows × 14 columns

9. Encoding Parameter

In [30]:

```
pd.read_csv('zomato.csv')
```

```
---------------------------------------------------------------------------
UnicodeDecodeError                        Traceback (most recent call last)
Cell In[30], line 1
----> 1 pd.read_csv('zomato.csv')

File ~\anaconda3\Lib\site-packages\pandas\util\_decorators.py:211, in deprecate_kwarg.<locals>._deprecate_kw
arg.<locals>.wrapper(*args, **kwargs)
    209     else:
    210         kwargs[new_arg_name] = new_arg_value
--> 211 return func(*args, **kwargs)

File ~\anaconda3\Lib\site-packages\pandas\util\_decorators.py:331, in deprecate_nonkeyword_arguments.<locals
>.decorate.<locals>.wrapper(*args, **kwargs)
    325 if len(args) > num_allow_args:
    326     warnings.warn(
    327         msg.format(arguments=_format_argument_list(allow_args)),
    328         FutureWarning,
    329         stacklevel=find_stack_level(),
    330     )
--> 331 return func(*args, **kwargs)

File ~\anaconda3\Lib\site-packages\pandas\io\parsers\readers.py:950, in read_csv(filepath_or_buffer, sep, de
limiter, header, names, index_col, usecols, squeeze, prefix, mangle_dupe_cols, dtype, engine, converters, tr
ue_values, false_values, skipinitialspace, skiprows, skipfooter, nrows, na_values, keep_default_na, na_filte
r, verbose, skip_blank_lines, parse_dates, infer_datetime_format, keep_date_col, date_parser, dayfirst, cach
e_dates, iterator, chunksize, compression, thousands, decimal, lineterminator, quotechar, quoting, doublequo
te, escapechar, comment, encoding, encoding_errors, dialect, error_bad_lines, warn_bad_lines, on_bad_lines,
delim_whitespace, low_memory, memory_map, float_precision, storage_options)
    935 kwds_defaults = _refine_defaults_read(
    936     dialect,
    937     delimiter,
   (...)
    946     defaults={"delimiter": ","},
    947 )
    948 kwds.update(kwds_defaults)
--> 950 return _read(filepath_or_buffer, kwds)

File ~\anaconda3\Lib\site-packages\pandas\io\parsers\readers.py:605, in _read(filepath_or_buffer, kwds)
    602 _validate_names(kwds.get("names", None))
    604 # Create the parser.
--> 605 parser = TextFileReader(filepath_or_buffer, **kwds)
    607 if chunksize or iterator:
    608     return parser

File ~\anaconda3\Lib\site-packages\pandas\io\parsers\readers.py:1442, in TextFileReader.__init__(self, f, en
gine, **kwds)
   1439     self.options["has_index_names"] = kwds["has_index_names"]
   1441 self.handles: IOHandles | None = None
-> 1442 self._engine = self._make_engine(f, self.engine)

File ~\anaconda3\Lib\site-packages\pandas\io\parsers\readers.py:1753, in TextFileReader._make_engine(self,
f, engine)
   1750     raise ValueError(msg)
   1752 try:
-> 1753     return mapping[engine](f, **self.options)
   1754 except Exception:
   1755     if self.handles is not None:

File ~\anaconda3\Lib\site-packages\pandas\io\parsers\c_parser_wrapper.py:79, in CParserWrapper.__init__(sel
f, src, **kwds)
     76     kwds.pop(key, None)
     78 kwds["dtype"] = ensure_dtype_objs(kwds.get("dtype", None))
---> 79 self._reader = parsers.TextReader(src, **kwds)
     81 self.unnamed_cols = self._reader.unnamed_cols
     83 # error: Cannot determine type of 'names'

File ~\anaconda3\Lib\site-packages\pandas\_libs\parsers.pyx:547, in pandas._libs.parsers.TextReader.__cinit_
_()

File ~\anaconda3\Lib\site-packages\pandas\_libs\parsers.pyx:636, in pandas._libs.parsers.TextReader._get_hea
der()

File ~\anaconda3\Lib\site-packages\pandas\_libs\parsers.pyx:852, in pandas._libs.parsers.TextReader._tokeniz
e_rows()

File ~\anaconda3\Lib\site-packages\pandas\_libs\parsers.pyx:1965, in pandas._libs.parsers.raise_parser_error
()

UnicodeDecodeError: 'utf-8' codec can't decode byte 0xed in position 7044: invalid continuation byte
```

In [31]:

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

Out[31]:

	Restaurant ID	Restaurant Name	Country Code	City	Address	Locality	Locality Verbose	Longitude	Latitude	Cuisines	...	C
0	6317637	Le Petit Souffle	162	Makati City	Third Floor, Century City Mall, Kalayaan Avenu...	Century City Mall, Poblacion, Makati City	Century City Mall, Poblacion, Makati City, Mak...	121.027535	14.565443	French, Japanese, Desserts	...	E
1	6304287	Izakaya Kikufuji	162	Makati City	Little Tokyo, 2277 Chino Roces Avenue, Legaspi...	Little Tokyo, Legaspi Village, Makati City	Little Tokyo, Legaspi Village, Makati City, Ma...	121.014101	14.553708	Japanese	...	E
2	6300002	Heat - Edsa Shangri-La	162	Mandaluyong City	Edsa Shangri-La, 1 Garden Way, Ortigas, Mandal...	Edsa Shangri-La, Ortigas, Mandaluyong City	Edsa Shangri-La, Ortigas, Mandaluyong City, Ma...	121.056831	14.581404	Seafood, Asian, Filipino, Indian	...	E
3	6318506	Ooma	162	Mandaluyong City	Third Floor, Mega Fashion Hall, SM Megamall, O...	SM Megamall, Ortigas, Mandaluyong City	SM Megamall, Ortigas, Mandaluyong City, Mandal...	121.056475	14.585318	Japanese, Sushi	...	E
4	6314302	Sambo Kojin	162	Mandaluyong City	Third Floor, Mega Atrium, SM Megamall, Ortigas...	SM Megamall, Ortigas, Mandaluyong City	SM Megamall, Ortigas, Mandaluyong City, Mandal...	121.057508	14.584450	Japanese, Korean	...	E
...	...	...	...	...	...	...	...	...	...	...	...	
9546	5915730	NamlÛ± Gurme	208	ÛÁstanbul	Kemanke□ô Karamustafa Pa□ôa Mahallesi, RÛ±htÛ±...	Karakí_y	Karakí_y, ÛÁstanbul	28.977392	41.022793	Turkish	...	
9547	5908749	Ceviz AÛôacÛ±	208	ÛÁstanbul	Ko□ôuyolu Mahallesi, Muhittin îstí_ndaÛô Cadd...	Ko□ôuyolu	Ko□ôuyolu, ÛÁstanbul	29.041297	41.009847	World Cuisine, Patisserie, Cafe	...	
9548	5915807	Huqqa	208	ÛÁstanbul	Kuruí_e□ôme Mahallesi, Muallim Naci Caddesi, N...	Kuruí_e□ôme	Kuruí_e□ôme, ÛÁstanbul	29.034640	41.055817	Italian, World Cuisine	...	
9549	5916112	A□ô□ôk Kahve	208	ÛÁstanbul	Kuruí_e□ôme Mahallesi, Muallim Naci Caddesi, N...	Kuruí_e□ôme	Kuruí_e□ôme, ÛÁstanbul	29.036019	41.057979	Restaurant Cafe	...	
9550	5927402	Walter's Coffee Roastery	208	ÛÁstanbul	CafeaÛôa Mahallesi, BademaltÛ± Sokak, No 21/B,...	Moda	Moda, ÛÁstanbul	29.026016	40.984776	Cafe	...	

9551 rows × 21 columns

10. Skip Bad Lines

In [32]:

```
pd.read_csv("BX-Books.csv")
```

```
---------------------------------------------------------------------------
UnicodeDecodeError                        Traceback (most recent call last)
Cell In[32], line 1
----> 1 pd.read_csv("BX-Books.csv")

File ~\anaconda3\Lib\site-packages\pandas\util\_decorators.py:211, in deprecate_kwarg.<locals>._deprecate_kw
arg.<locals>.wrapper(*args, **kwargs)
    209     else:
    210         kwargs[new_arg_name] = new_arg_value
--> 211 return func(*args, **kwargs)

File ~\anaconda3\Lib\site-packages\pandas\util\_decorators.py:331, in deprecate_nonkeyword_arguments.<locals
>.decorate.<locals>.wrapper(*args, **kwargs)
    325 if len(args) > num_allow_args:
    326     warnings.warn(
    327         msg.format(arguments=_format_argument_list(allow_args)),
    328         FutureWarning,
    329         stacklevel=find_stack_level(),
    330     )
--> 331 return func(*args, **kwargs)

File ~\anaconda3\Lib\site-packages\pandas\io\parsers\readers.py:950, in read_csv(filepath_or_buffer, sep, de
limiter, header, names, index_col, usecols, squeeze, prefix, mangle_dupe_cols, dtype, engine, converters, tr
ue_values, false_values, skipinitialspace, skiprows, skipfooter, nrows, na_values, keep_default_na, na_filte
r, verbose, skip_blank_lines, parse_dates, infer_datetime_format, keep_date_col, date_parser, dayfirst, cach
e_dates, iterator, chunksize, compression, thousands, decimal, lineterminator, quotechar, quoting, doublequo
te, escapechar, comment, encoding, encoding_errors, dialect, error_bad_lines, warn_bad_lines, on_bad_lines,
delim_whitespace, low_memory, memory_map, float_precision, storage_options)
    935 kwds_defaults = _refine_defaults_read(
    936     dialect,
    937     delimiter,
   (...)
    946     defaults={"delimiter": ","},
    947 )
    948 kwds.update(kwds_defaults)
--> 950 return _read(filepath_or_buffer, kwds)

File ~\anaconda3\Lib\site-packages\pandas\io\parsers\readers.py:605, in _read(filepath_or_buffer, kwds)
    602 _validate_names(kwds.get("names", None))
    604 # Create the parser.
--> 605 parser = TextFileReader(filepath_or_buffer, **kwds)
    607 if chunksize or iterator:
    608     return parser

File ~\anaconda3\Lib\site-packages\pandas\io\parsers\readers.py:1442, in TextFileReader.__init__(self, f, en
gine, **kwds)
   1439     self.options["has_index_names"] = kwds["has_index_names"]
   1441 self.handles: IOHandles | None = None
-> 1442 self._engine = self._make_engine(f, self.engine)

File ~\anaconda3\Lib\site-packages\pandas\io\parsers\readers.py:1753, in TextFileReader._make_engine(self,
f, engine)
   1750     raise ValueError(msg)
   1752 try:
-> 1753     return mapping[engine](f, **self.options)
   1754 except Exception:
   1755     if self.handles is not None:

File ~\anaconda3\Lib\site-packages\pandas\io\parsers\c_parser_wrapper.py:79, in CParserWrapper.__init__(sel
f, src, **kwds)
     76     kwds.pop(key, None)
     78 kwds["dtype"] = ensure_dtype_objs(kwds.get("dtype", None))
---> 79 self._reader = parsers.TextReader(src, **kwds)
     81 self.unnamed_cols = self._reader.unnamed_cols
     83 # error: Cannot determine type of 'names'

File ~\anaconda3\Lib\site-packages\pandas\_libs\parsers.pyx:547, in pandas._libs.parsers.TextReader.__cinit_
_()

File ~\anaconda3\Lib\site-packages\pandas\_libs\parsers.pyx:636, in pandas._libs.parsers.TextReader._get_hea
der()

File ~\anaconda3\Lib\site-packages\pandas\_libs\parsers.pyx:852, in pandas._libs.parsers.TextReader._tokeniz
e_rows()

File ~\anaconda3\Lib\site-packages\pandas\_libs\parsers.pyx:1965, in pandas._libs.parsers.raise_parser_error
()

UnicodeDecodeError: 'utf-8' codec can't decode byte 0xc3 in position 9431: invalid continuation byte
```

In [36]:

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

```
C:\Users\TOSHIBA\AppData\Local\Temp\ipykernel_1544\115290075.py:1: FutureWarning: The error_bad_lines argume
nt has been deprecated and will be removed in a future version. Use on_bad_lines in the future.


  pd.read_csv('BX-Books.csv', sep=';', encoding='latin-1', error_bad_lines=False)
Skipping line 6452: expected 8 fields, saw 9
Skipping line 43667: expected 8 fields, saw 10
Skipping line 51751: expected 8 fields, saw 9

Skipping line 92038: expected 8 fields, saw 9
Skipping line 104319: expected 8 fields, saw 9
Skipping line 121768: expected 8 fields, saw 9

Skipping line 144058: expected 8 fields, saw 9
Skipping line 150789: expected 8 fields, saw 9
Skipping line 157128: expected 8 fields, saw 9
Skipping line 180189: expected 8 fields, saw 9
Skipping line 185738: expected 8 fields, saw 9

Skipping line 209388: expected 8 fields, saw 9
Skipping line 220626: expected 8 fields, saw 9
Skipping line 227933: expected 8 fields, saw 11
Skipping line 228957: expected 8 fields, saw 10
Skipping line 245933: expected 8 fields, saw 9
Skipping line 251296: expected 8 fields, saw 9
Skipping line 259941: expected 8 fields, saw 9
Skipping line 261529: expected 8 fields, saw 9

C:\Users\TOSHIBA\AppData\Local\Temp\ipykernel_1544\115290075.py:1: DtypeWarning: Columns (3) have mixed type
s. Specify dtype option on import or set low_memory=False.
  pd.read_csv('BX-Books.csv', sep=';', encoding='latin-1', error_bad_lines=False)
```

Out[36]:

	ISBN	Book-Title	Book-Author	Year-Of-Publication	Publisher	Image-URL-S	
0	0195153448	Classical Mythology	Mark P. O. Morford	2002	Oxford University Press	http://images.amazon.com/images/P/0195153448.0...	http://imag
1	0002005018	Clara Callan	Richard Bruce Wright	2001	HarperFlamingo Canada	http://images.amazon.com/images/P/0002005018.0...	http://imag
2	0060973129	Decision in Normandy	Carlo D'Este	1991	HarperPerennial	http://images.amazon.com/images/P/0060973129.0...	http://imag
3	0374157065	Flu: The Story of the Great Influenza Pandemic...	Gina Bari Kolata	1999	Farrar Straus Giroux	http://images.amazon.com/images/P/0374157065.0...	http://imag
4	0393045218	The Mummies of Urumchi	E. J. W. Barber	1999	W. W. Norton & Company	http://images.amazon.com/images/P/0393045218.0...	http://imag
...	...	...	...	...	...	...	
271355	0440400988	There's a Bat in Bunk Five	Paula Danziger	1988	Random House Childrens Pub (Mm)	http://images.amazon.com/images/P/0440400988.0...	http://imag
271356	0525447644	From One to One Hundred	Teri Sloat	1991	Dutton Books	http://images.amazon.com/images/P/0525447644.0...	http://imag
271357	006008667X	Lily Dale : The True Story of the Town that Ta...	Christine Wicker	2004	HarperSanFrancisco	http://images.amazon.com/images/P/006008667X.0...	http://imag
271358	0192126040	Republic (World's Classics)	Plato	1996	Oxford University Press	http://images.amazon.com/images/P/0192126040.0...	http://imag
271359	0767409752	A Guided Tour of Rene Descartes' Meditations o...	Christopher Biffle	2000	McGraw-Hill Humanities/Social Sciences/Languages	http://images.amazon.com/images/P/0767409752.0...	http://imag

271360 rows × 8 columns

11.dtypes parameter

In [41]:

```
pd.read_csv('aug_train.csv', dtype={'target' : int}).info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 19158 entries, 0 to 19157
Data columns (total 14 columns):
 #   Column                  Non-Null Count  Dtype  
---  ------                  --------------  -----  
 0   enrollee_id             19158 non-null  int64  
 1   city                    19158 non-null  object 
 2   city_development_index  19158 non-null  float64
 3   gender                  14650 non-null  object 
 4   relevent_experience     19158 non-null  object 
 5   enrolled_university     18772 non-null  object 
 6   education_level         18698 non-null  object 
 7   major_discipline        16345 non-null  object 
 8   experience              19093 non-null  object 
 9   company_size            13220 non-null  object 
 10  company_type            13018 non-null  object 
 11  last_new_job            18735 non-null  object 
 12  training_hours          19158 non-null  int64  
 13  target                  19158 non-null  int32  
dtypes: float64(1), int32(1), int64(2), object(10)
memory usage: 2.0+ MB
```

12. Loading a Huge dataset in chunks

In [48]:

```
dfs = pd.read_csv('aug_train.csv', chunksize=5000)
```

In [50]:

```
for chunks in dfs:
    print(chunks.shape)
```

```
(5000, 14)
(5000, 14)
(4158, 14)
```

In [ ]:

```

```