```
#importing librarie
import pandas as pd
import numpy as np


#reading the csv file
df = pd.read_csv("open_pubs.csv")


#viewing first five records
df.head()
```

	22	Anchor Inn	Upper Street, Stratford St Mary, COLCHESTER	CO7 6LW	604749	234404	51.970379	0.979340	Babergh
0	36	Ark Bar Restaurant	Ark Bar And Restaurant, Cattawade Street, Bran...	CO11 1RH	610194	233329	51.958698	1.057832	Babergh
1	74	Black Boy	The Lady Elizabeth, 7 Market Hill, SUDBURY, Su...	CO10 2EA	587334	241316	52.038595	0.729915	Babergh

```
#reloading csv file
df = pd.read_csv("open_pubs.csv",header=None)


df.head()
```

	0	1	2	3	4	5	6	7	8
0	22	Anchor Inn	Upper Street, Stratford St Mary, COLCHESTER	CO7 6LW	604749	234404	51.970379	0.979340	Babergh
1	36	Ark Bar Restaurant	Ark Bar And Restaurant, Cattawade Street, Bran...	CO11 1RH	610194	233329	51.958698	1.057832	Babergh
			The Lady Elizabeth, 7						

```
#naming the headers
df.columns=['fsa_id','name','address','postcode','easting','northing','latitude','longitude','local_authority']


df.head()
```

	fsa_id	name	address	postcode	easting	northing	latitude	longitude	local_
0	22	Anchor Inn	Upper Street, Stratford St Mary, COLCHESTER	CO7 6LW	604749	234404	51.970379	0.979340	
1	36	Ark Bar Restaurant	Ark Bar And Restaurant, Cattawade Street, Bran...	CO11 1RH	610194	233329	51.958698	1.057832	

```
#viewing number of rows and columns in dataset
df.shape
```

```
(51331, 9)
```

```
#information regarding the dataset
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 51331 entries, 0 to 51330
Data columns (total 9 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   fsa_id           51331 non-null  int64
 1   name             51331 non-null  object
 2   address          51331 non-null  object
 3   postcode         51331 non-null  object
 4   easting          51331 non-null  int64
```

```
 5   northing         51331 non-null  int64
 6   latitude         51331 non-null  object
 7   longitude        51331 non-null  object
 8   local_authority  51331 non-null  object
dtypes: int64(3), object(6)
memory usage: 3.5+ MB
```

```
#viewing the null values
df.isnull().sum()
```

```
fsa_id             0
name               0
address            0
postcode           0
easting            0
northing           0
latitude           0
longitude          0
local_authority    0
dtype: int64
```

```
#checking for duplicate rows
df.duplicated().sum()
```

```
0
```

```
#printing unique values of each column
print("Printing unique values of each column")
for i in df.columns[:]:
  print("*"*70)
  print("number of unique values of ",i," column :",df[i].nunique())
  print("unique values:")
  print(df[i].unique())
```

```
number of unique values of  address  column : 50013
unique values:
['Upper Street, Stratford St Mary, COLCHESTER'
 'Ark Bar And Restaurant, Cattawade Street, Brantham, MANNINGTREE'
 'The Lady Elizabeth, 7 Market Hill, SUDBURY, Suffolk' ...
 'Sports And Social Club, 44 Brook Street, Wrexham, Wrexham'
 'Wrexham Rugby Club Bryn Estyn Road, Wrexham, Wrexham'
 'Y Tai Railway Road, Brymbo, Wrexham, Wrexham']
**********************************************************************
number of unique values of  postcode  column : 45886
unique values:
['CO7 6LW' 'CO11 1RH' 'CO10 2EA' ... 'LL13 7LU' 'LL13 9TY' 'LL11 5EA']
**********************************************************************
number of unique values of  easting  column : 43229
unique values:
[604749 610194 587334 ... 333259 335808 329721]
**********************************************************************
number of unique values of  northing  column : 43896
unique values:
[234404 233329 241316 ... 350213 351078 353785]
**********************************************************************
number of unique values of  latitude  column : 46810
unique values:
['51.970379' '51.958698' '52.038595' ... '53.044998' '53.053094'
 '53.076638']
**********************************************************************
number of unique values of  longitude  column : 46807
unique values:
['0.979340' '1.057832' '0.729915' ... '-2.996966' '-2.959124' '-3.050512']
**********************************************************************
number of unique values of  local_authority  column : 360
unique values:
['Babergh' 'Mid Suffolk' 'Basildon' 'Bedford' 'Huntingdonshire'
 'Braintree' 'Breckland' 'Brentwood' 'Broadland' 'Broxbourne' 'Cambridge'
 'Castle Point' 'Central Bedfordshire' 'Luton' 'Dacorum' 'Chelmsford'
 'Rochford' 'Colchester' 'East Cambridgeshire' 'East Hertfordshire'
```

```
'Oadby and Wigston' 'Rushcliffe' 'Rutland' 'South Derbyshire'
'East Staffordshire' 'South Holland' 'South Kesteven' 'West Lindsey'
'West Northamptonshire' 'Barking and Dagenham' 'Barnet' 'Bexley' 'Brent'
'Harrow' 'Bromley' 'Croydon' 'Camden' 'City of London' 'Lambeth' 'Ealing'
'Hillingdon' 'Enfield' 'Waltham Forest' 'Greenwich' 'Hackney'
'Tower Hamlets' 'Hammersmith and Fulham' 'Haringey' 'Havering' 'Hounslow'
'Islington' 'Kensington and Chelsea' 'Kingston upon Thames' 'Lewisham'
'Merton' 'Newham' 'Richmond upon Thames' 'Wandsworth' 'Southwark'
'Sutton' 'Westminster' 'Darlington' 'County Durham' 'Gateshead'
'Hartlepool' 'Middlesbrough' 'Newcastle upon Tyne' 'North Tyneside'
'Northumberland' 'Redcar and Cleveland' 'South Tyneside'
'Stockton-on-Tees' 'Sunderland' 'Allerdale' 'Barrow-in-Furness'
```

```
df.latitude.value_counts()
```

```
\N          767
51.310946    10
52.896164    10
53.178189     9
55.318954     8
            ...
53.369771     1
53.390799     1
53.371238     1
53.361078     1
53.076638     1
Name: latitude, Length: 46810, dtype: int64
```

```
df.longitude.value_counts()
```

```
\N          767
-0.655576    12
-1.853894    10
0.349896      9
-2.051721     8
            ...
-2.734245     1
-2.735554     1
-2.733804     1
-2.735118     1
-3.050512     1
Name: longitude, Length: 46807, dtype: int64
```

```
#replacing data with NaN
df['latitude'] = df['latitude'].replace(['\\N'],np.NaN)
df['longitude'] = df['longitude'].replace(['\\N'],np.NaN)
```

```
df.isna().sum()
```

```
fsa_id              0
name                0
address             0
postcode            0
easting             0
northing            0
latitude          767
longitude         767
local_authority     0
dtype: int64
```

```
df.shape
```

```
(51331, 9)
```

```
#dropping the rows with null value
df.dropna(inplace=True)
```

```
df.isna().sum()
```

```
fsa_id            0
name              0
address           0
postcode          0
easting           0
northing          0
latitude          0
longitude         0
```

```
local_authority    0
dtype: int64
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 50564 entries, 0 to 51330
Data columns (total 9 columns):
 #   Column           Non-Null Count  Dtype 
---  ------           --------------  ----- 
 0   fsa_id           50564 non-null  int64 
 1   name             50564 non-null  object
 2   address          50564 non-null  object
 3   postcode         50564 non-null  object
 4   easting          50564 non-null  int64 
 5   northing         50564 non-null  int64 
 6   latitude         50564 non-null  object
 7   longitude        50564 non-null  object
 8   local_authority  50564 non-null  object
dtypes: int64(3), object(6)
memory usage: 3.9+ MB
```

```
df.sample(5)
```

	fsa_id	name	address	postcode	easting	northing	latitude	longitude	local_aut
33516	388887	Codford Village Hall & Bar	Codford Sports And Social Club, Codford Villag...	BA12 0PP	396898	139799	51.157426	-2.045733	V
			2b						

```
# Converting categorical values to numerical values

df.latitude = df.latitude.astype(float)
df.longitude = df.longitude.astype(float)
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 50564 entries, 0 to 51330
Data columns (total 9 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   fsa_id           50564 non-null  int64  
 1   name             50564 non-null  object 
 2   address          50564 non-null  object 
 3   postcode         50564 non-null  object 
 4   easting          50564 non-null  int64  
 5   northing         50564 non-null  int64  
 6   latitude         50564 non-null  float64
 7   longitude        50564 non-null  float64
 8   local_authority  50564 non-null  object 
dtypes: float64(2), int64(3), object(4)
memory usage: 3.9+ MB
```

```
df.head()
```

	fsa_id	name	address	postcode	easting	northing	latitude	longitude	local_
0	22	Anchor Inn	Upper Street, Stratford St Mary, COLCHESTER	CO7 6LW	604749	234404	51.970379	0.979340	
1	36	Ark Bar Restaurant	Ark Bar And Restaurant, Cattawade Street, Bran...	CO11 1RH	610194	233329	51.958698	1.057832	

```
df.shape
```

```
(50564, 9)
```

```
df.to_csv('clean_open_pubs.csv',index=False)
```

```
from google.colab import files
df.to_csv('clean_open_pubs.csv', index=False,encoding = 'utf-8')
files.download('clean_open_pubs.csv')
```