datacleaning-file-1

November 1, 2024

[1]:

```
import pandas as pd

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

[2]:

```
print(df)
```

```
      Loan_ID  Gender Married Dependents     Education Self_Employed  \
0    LP001002    Male      No          0      Graduate            No
1    LP001003    Male     Yes          1      Graduate            No
2    LP001005    Male     Yes          0      Graduate           Yes
3    LP001006    Male     Yes          0  Not Graduate            No
4    LP001008    Male      No          0      Graduate            No
..        …       …     …        …           …           …
609  LP002978  Female      No          0      Graduate            No
610  LP002979    Male     Yes         3+      Graduate            No
611  LP002983    Male     Yes          1      Graduate            No
612  LP002984    Male     Yes          2      Graduate            No
613  LP002990  Female      No          0      Graduate           Yes

     ApplicantIncome  CoapplicantIncome  LoanAmount  Loan_Amount_Term  \
0               5849                0.0         NaN             360.0
1               4583             1508.0       128.0             360.0
2               3000                0.0        66.0             360.0
3               2583             2358.0       120.0             360.0
4               6000                0.0       141.0             360.0
..               …                …         …               …
609             2900                0.0        71.0             360.0
610             4106                0.0        40.0             180.0
611             8072              240.0       253.0             360.0
612             7583                0.0       187.0             360.0
613             4583                0.0       133.0             360.0

     Credit_History Property_Area Loan_Status
0               1.0         Urban           Y
1               1.0         Rural           N
2               1.0         Urban           Y
3               1.0         Urban           Y
4               1.0         Urban           Y
```

```
..              …           …         …
609             1.0         Rural               Y
610             1.0         Rural               Y
611             1.0         Urban               Y
612             1.0         Urban               Y
613             0.0     Semiurban               N

[614 rows x 13 columns]
```

```
[3]: df.drop(['Dependents'], axis=1) #drop the column
```

```
[3]:       Loan_ID  Gender Married     Education Self_Employed  ApplicantIncome  \
     0    LP001002    Male      No      Graduate            No             5849
     1    LP001003    Male     Yes      Graduate            No             4583
     2    LP001005    Male     Yes      Graduate           Yes             3000
     3    LP001006    Male     Yes  Not Graduate            No             2583
     4    LP001008    Male      No      Graduate            No             6000
     ..        …     …      …           …           …              …
     609  LP002978  Female      No      Graduate            No             2900
     610  LP002979    Male     Yes      Graduate            No             4106
     611  LP002983    Male     Yes      Graduate            No             8072
     612  LP002984    Male     Yes      Graduate            No             7583
     613  LP002990  Female      No      Graduate           Yes             4583

          CoapplicantIncome  LoanAmount  Loan_Amount_Term  Credit_History  \
     0                  0.0         NaN             360.0             1.0
     1               1508.0       128.0             360.0             1.0
     2                  0.0        66.0             360.0             1.0
     3               2358.0       120.0             360.0             1.0
     4                  0.0       141.0             360.0             1.0
     ..                 …         …               …             …
     609                0.0        71.0             360.0             1.0
     610                0.0        40.0             180.0             1.0
     611              240.0       253.0             360.0             1.0
     612                0.0       187.0             360.0             1.0
     613                0.0       133.0             360.0             0.0

         Property_Area Loan_Status
     0           Urban           Y
     1           Rural           N
     2           Urban           Y
     3           Urban           Y
     4           Urban           Y
     ..            …         …
     609         Rural           Y
     610         Rural           Y
     611         Urban           Y
```

```
612         Urban           Y
613     Semiurban           N

[614 rows x 12 columns]
```

[4]:

```
df.drop([0, 1])   #drop the rows
```

[4]:

```
      Loan_ID  Gender Married Dependents     Education Self_Employed  \
2    LP001005    Male     Yes          0      Graduate           Yes
3    LP001006    Male     Yes          0  Not Graduate            No
4    LP001008    Male      No          0      Graduate            No
5    LP001011    Male     Yes          2      Graduate           Yes
6    LP001013    Male     Yes          0  Not Graduate            No
..        ...     ...     ...        ...           ...           ...
609  LP002978  Female      No          0      Graduate            No
610  LP002979    Male     Yes         3+      Graduate            No
611  LP002983    Male     Yes          1      Graduate            No
612  LP002984    Male     Yes          2      Graduate            No
613  LP002990  Female      No          0      Graduate           Yes

     ApplicantIncome  CoapplicantIncome  LoanAmount  Loan_Amount_Term  \
2               3000                0.0        66.0             360.0
3               2583             2358.0       120.0             360.0
4               6000                0.0       141.0             360.0
5               5417             4196.0       267.0             360.0
6               2333             1516.0        95.0             360.0
..               ...                ...         ...               ...
609             2900                0.0        71.0             360.0
610             4106                0.0        40.0             180.0
611             8072              240.0       253.0             360.0
612             7583                0.0       187.0             360.0
613             4583                0.0       133.0             360.0

     Credit_History Property_Area Loan_Status
2               1.0         Urban           Y
3               1.0         Urban           Y
4               1.0         Urban           Y
5               1.0         Urban           Y
6               1.0         Urban           Y
..              ...           ...         ...
609             1.0         Rural           Y
610             1.0         Rural           Y
611             1.0         Urban           Y
612             1.0         Urban           Y
613             0.0     Semiurban           N

[612 rows x 13 columns]
```

[5]:

```
df.columns[0]  #displays 1st column name
```

[5]:

```
'Loan_ID'
```

[6]:

```
import pandas as pd
import numpy as np
df = pd.read_csv("train.csv")
print(df.replace(np.NaN,0))
#df['DataFrame Column'] = df['DataFrame Column'].replace(np.nan, 0)
```

```
      Loan_ID  Gender Married Dependents     Education Self_Employed  \
0    LP001002    Male      No          0      Graduate            No
1    LP001003    Male     Yes          1      Graduate            No
2    LP001005    Male     Yes          0      Graduate           Yes
3    LP001006    Male     Yes          0  Not Graduate            No
4    LP001008    Male      No          0      Graduate            No
..        …     …     …        …           …           …
609  LP002978  Female      No          0      Graduate            No
610  LP002979    Male     Yes         3+      Graduate            No
611  LP002983    Male     Yes          1      Graduate            No
612  LP002984    Male     Yes          2      Graduate            No
613  LP002990  Female      No          0      Graduate           Yes

     ApplicantIncome  CoapplicantIncome  LoanAmount  Loan_Amount_Term  \
0               5849                0.0         0.0             360.0
1               4583             1508.0       128.0             360.0
2               3000                0.0        66.0             360.0
3               2583             2358.0       120.0             360.0
4               6000                0.0       141.0             360.0
..                …                …         …               …
609             2900                0.0        71.0             360.0
610             4106                0.0        40.0             180.0
611             8072              240.0       253.0             360.0
612             7583                0.0       187.0             360.0
613             4583                0.0       133.0             360.0

     Credit_History Property_Area Loan_Status
0               1.0         Urban           Y
1               1.0         Rural           N
2               1.0         Urban           Y
3               1.0         Urban           Y
4               1.0         Urban           Y
..              …           …         …
609             1.0         Rural           Y
610             1.0         Rural           Y
611             1.0         Urban           Y
612             1.0         Urban           Y
```

```
613             0.0     Semiurban           N

[614 rows x 13 columns]
```

[7]:

```python
print ("NaN replaced with '0':")
print( df.fillna(method='pad'))
```

```
NaN replaced with '0':
      Loan_ID  Gender Married Dependents     Education Self_Employed   \
0    LP001002    Male      No          0      Graduate            No
1    LP001003    Male     Yes          1      Graduate            No
2    LP001005    Male     Yes          0      Graduate           Yes
3    LP001006    Male     Yes          0  Not Graduate            No
4    LP001008    Male      No          0      Graduate            No
..        …     …     …        …           …           …
609  LP002978  Female      No          0      Graduate            No
610  LP002979    Male     Yes         3+      Graduate            No
611  LP002983    Male     Yes          1      Graduate            No
612  LP002984    Male     Yes          2      Graduate            No
613  LP002990  Female      No          0      Graduate           Yes

     ApplicantIncome  CoapplicantIncome  LoanAmount  Loan_Amount_Term  \
0               5849                0.0         NaN             360.0
1               4583             1508.0       128.0             360.0
2               3000                0.0        66.0             360.0
3               2583             2358.0       120.0             360.0
4               6000                0.0       141.0             360.0
..               …                …         …               …
609             2900                0.0        71.0             360.0
610             4106                0.0        40.0             180.0
611             8072              240.0       253.0             360.0
612             7583                0.0       187.0             360.0
613             4583                0.0       133.0             360.0

     Credit_History Property_Area Loan_Status
0               1.0         Urban           Y
1               1.0         Rural           N
2               1.0         Urban           Y
3               1.0         Urban           Y
4               1.0         Urban           Y
..              …           …         …
609             1.0         Rural           Y
610             1.0         Rural           Y
611             1.0         Urban           Y
612             1.0         Urban           Y
613             0.0     Semiurban           N

[614 rows x 13 columns]
```

```
C:\Users\Prateek\AppData\Local\Temp\ipykernel_17908\2002809902.py:2:
FutureWarning: DataFrame.fillna with 'method' is deprecated and will raise in a
future version. Use obj.ffill() or obj.bfill() instead.
  print( df.fillna(method='pad'))
```

```
[8]: print (df['Loan_ID'].isnull())
```

```
0      False
1      False
2      False
3      False
4      False
       …
609    False
610    False
611    False
612    False
613    False
Name: Loan_ID, Length: 614, dtype: bool
```

```
[9]: print (df['Dependents'].notnull())
```

```
0      True
1      True
2      True
3      True
4      True
       …
609    True
610    True
611    True
612    True
613    True
Name: Dependents, Length: 614, dtype: bool
```

```
[10]: print( df['Self_Employed'].isnull())
```

```
0      False
1      False
2      False
3      False
4      False
       …
609    False
610    False
611    False
612    False
613    False
```

```
Name: Self_Employed, Length: 614, dtype: bool
```

[11]:
```
print(df)
print ("NaN replaced with '0':")
print( df.fillna(0))
```

```
      Loan_ID  Gender Married Dependents     Education Self_Employed  \
0    LP001002    Male      No          0      Graduate            No   
1    LP001003    Male     Yes          1      Graduate            No   
2    LP001005    Male     Yes          0      Graduate           Yes   
3    LP001006    Male     Yes          0  Not Graduate            No   
4    LP001008    Male      No          0      Graduate            No   
..        …       …       …          …             …             …   
609  LP002978  Female      No          0      Graduate            No   
610  LP002979    Male     Yes         3+      Graduate            No   
611  LP002983    Male     Yes          1      Graduate            No   
612  LP002984    Male     Yes          2      Graduate            No   
613  LP002990  Female      No          0      Graduate           Yes   

     ApplicantIncome  CoapplicantIncome  LoanAmount  Loan_Amount_Term  \
0               5849                0.0         NaN             360.0   
1               4583             1508.0       128.0             360.0   
2               3000                0.0        66.0             360.0   
3               2583             2358.0       120.0             360.0   
4               6000                0.0       141.0             360.0   
..               …                …         …               …   
609             2900                0.0        71.0             360.0   
610             4106                0.0        40.0             180.0   
611             8072              240.0       253.0             360.0   
612             7583                0.0       187.0             360.0   
613             4583                0.0       133.0             360.0   

     Credit_History Property_Area Loan_Status  
0               1.0         Urban           Y  
1               1.0         Rural           N  
2               1.0         Urban           Y  
3               1.0         Urban           Y  
4               1.0         Urban           Y  
..              …           …         …  
609             1.0         Rural           Y  
610             1.0         Rural           Y  
611             1.0         Urban           Y  
612             1.0         Urban           Y  
613             0.0     Semiurban           N  

[614 rows x 13 columns]
NaN replaced with '0':
      Loan_ID  Gender Married Dependents     Education Self_Employed  \
0    LP001002    Male      No          0      Graduate            No
1    LP001003    Male     Yes          1      Graduate            No
2    LP001005    Male     Yes          0      Graduate           Yes
3    LP001006    Male     Yes          0  Not Graduate            No
4    LP001008    Male      No          0      Graduate            No
..        …     …     …        …           …           …
609  LP002978  Female      No          0      Graduate            No
610  LP002979    Male     Yes         3+      Graduate            No
611  LP002983    Male     Yes          1      Graduate            No
612  LP002984    Male     Yes          2      Graduate            No
613  LP002990  Female      No          0      Graduate           Yes

     ApplicantIncome  CoapplicantIncome  LoanAmount  Loan_Amount_Term  \
0               5849                0.0         0.0             360.0
1               4583             1508.0       128.0             360.0
2               3000                0.0        66.0             360.0
3               2583             2358.0       120.0             360.0
4               6000                0.0       141.0             360.0
..               …                …         …               …
609             2900                0.0        71.0             360.0
610             4106                0.0        40.0             180.0
611             8072              240.0       253.0             360.0
612             7583                0.0       187.0             360.0
613             4583                0.0       133.0             360.0

     Credit_History Property_Area Loan_Status
0               1.0         Urban           Y
1               1.0         Rural           N
2               1.0         Urban           Y
3               1.0         Urban           Y
4               1.0         Urban           Y
..              …           …         …
609             1.0         Rural           Y
610             1.0         Rural           Y
611             1.0         Urban           Y
612             1.0         Urban           Y
613             0.0     Semiurban           N

[614 rows x 13 columns]
```

[12]:
```
df = df.dropna() #drops rows with null values
```

[13]:
```
print(df)
```

```
      Loan_ID  Gender Married Dependents     Education Self_Employed  \
1    LP001003    Male     Yes          1      Graduate            No
2    LP001005    Male     Yes          0      Graduate           Yes
3    LP001006    Male     Yes          0  Not Graduate            No
```

```
4    LP001008    Male      No          0      Graduate            No
5    LP001011    Male     Yes          2      Graduate           Yes
..        …     …     …        …           …           …
609  LP002978  Female      No          0      Graduate            No
610  LP002979    Male     Yes         3+      Graduate            No
611  LP002983    Male     Yes          1      Graduate            No
612  LP002984    Male     Yes          2      Graduate            No
613  LP002990  Female      No          0      Graduate           Yes

     ApplicantIncome  CoapplicantIncome  LoanAmount  Loan_Amount_Term  \
1               4583             1508.0       128.0             360.0
2               3000                0.0        66.0             360.0
3               2583             2358.0       120.0             360.0
4               6000                0.0       141.0             360.0
5               5417             4196.0       267.0             360.0
..               …                …         …               …
609             2900                0.0        71.0             360.0
610             4106                0.0        40.0             180.0
611             8072              240.0       253.0             360.0
612             7583                0.0       187.0             360.0
613             4583                0.0       133.0             360.0

     Credit_History Property_Area Loan_Status
1               1.0         Rural           N
2               1.0         Urban           Y
3               1.0         Urban           Y
4               1.0         Urban           Y
5               1.0         Urban           Y
..              …           …         …
609             1.0         Rural           Y
610             1.0         Rural           Y
611             1.0         Urban           Y
612             1.0         Urban           Y
613             0.0     Semiurban           N

[480 rows x 13 columns]
```

[14]:

```
import pandas as pd
import numpy as np
df = pd.read_csv("train.csv")
df.head()
```

[14]:

```
    Loan_ID Gender Married Dependents     Education Self_Employed  \
0  LP001002   Male      No          0      Graduate            No
1  LP001003   Male     Yes          1      Graduate            No
2  LP001005   Male     Yes          0      Graduate           Yes
3  LP001006   Male     Yes          0  Not Graduate            No
```

```
4  LP001008   Male      No          0      Graduate            No

   ApplicantIncome  CoapplicantIncome  LoanAmount  Loan_Amount_Term  \
0             5849                0.0         NaN             360.0
1             4583             1508.0       128.0             360.0
2             3000                0.0        66.0             360.0
3             2583             2358.0       120.0             360.0
4             6000                0.0       141.0             360.0

   Credit_History Property_Area Loan_Status
0             1.0         Urban           Y
1             1.0         Rural           N
2             1.0         Urban           Y
3             1.0         Urban           Y
4             1.0         Urban           Y
```

[15]:

```
to_drop = ['Gender','Married']
#df.drop(columns=to_drop, inplace=True)
df.drop(to_drop, inplace=True, axis=1)
```

[16]:

```
df.head()
```

[16]:

```
    Loan_ID Dependents     Education Self_Employed  ApplicantIncome  \
0  LP001002          0      Graduate            No             5849
1  LP001003          1      Graduate            No             4583
2  LP001005          0      Graduate           Yes             3000
3  LP001006          0  Not Graduate            No             2583
4  LP001008          0      Graduate            No             6000

   CoapplicantIncome  LoanAmount  Loan_Amount_Term  Credit_History  \
0                0.0         NaN             360.0             1.0
1             1508.0       128.0             360.0             1.0
2                0.0        66.0             360.0             1.0
3             2358.0       120.0             360.0             1.0
4                0.0       141.0             360.0             1.0

  Property_Area Loan_Status
0         Urban           Y
1         Rural           N
2         Urban           Y
3         Urban           Y
4         Urban           Y
```

[17]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 614 entries, 0 to 613
```

```
Data columns (total 11 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Loan_ID            614 non-null    object
 1   Dependents         599 non-null    object
 2   Education          614 non-null    object
 3   Self_Employed      582 non-null    object
 4   ApplicantIncome    614 non-null    int64
 5   CoapplicantIncome  614 non-null    float64
 6   LoanAmount         592 non-null    float64
 7   Loan_Amount_Term   600 non-null    float64
 8   Credit_History     564 non-null    float64
 9   Property_Area      614 non-null    object
 10  Loan_Status        614 non-null    object
dtypes: float64(4), int64(1), object(6)
memory usage: 52.9+ KB
```

[18]:
```
df.shape
```

[18]:
```
(614, 11)
```

[19]:
```
df.count()
```

[19]:
```
Loan_ID              614
Dependents           599
Education            614
Self_Employed        582
ApplicantIncome      614
CoapplicantIncome    614
LoanAmount           592
Loan_Amount_Term     600
Credit_History       564
Property_Area        614
Loan_Status          614
dtype: int64
```

[20]:
```
df.isnull()
```

[20]:
```
     Loan_ID  Dependents  Education  Self_Employed  ApplicantIncome  \
0      False       False      False          False            False
1      False       False      False          False            False
2      False       False      False          False            False
3      False       False      False          False            False
4      False       False      False          False            False
..       …           …        …              …                …
609    False       False      False          False            False
610    False       False      False          False            False
```

```
611    False       False      False          False            False
612    False       False      False          False            False
613    False       False      False          False            False

     CoapplicantIncome  LoanAmount  Loan_Amount_Term  Credit_History  \
0                False        True             False           False
1                False       False             False           False
2                False       False             False           False
3                False       False             False           False
4                False       False             False           False
..                   …         …               …             …
609              False       False             False           False
610              False       False             False           False
611              False       False             False           False
612              False       False             False           False
613              False       False             False           False

     Property_Area  Loan_Status
0            False        False
1            False        False
2            False        False
3            False        False
4            False        False
..               …          …
609          False        False
610          False        False
611          False        False
612          False        False
613          False        False

[614 rows x 11 columns]
```

[21]:

```
missing_values=df.isnull()
```

[22]:

```
missing_values.dtypes
```

[22]:

```
Loan_ID              bool
Dependents           bool
Education            bool
Self_Employed        bool
ApplicantIncome      bool
CoapplicantIncome    bool
LoanAmount           bool
Loan_Amount_Term     bool
Credit_History       bool
Property_Area        bool
Loan_Status          bool
```

```
      dtype: object
```

```
[23]: no_missing_values=missing_values.sum()
```

```
[24]: missing_values.sum()
```

```
[24]: Loan_ID               0
      Dependents           15
      Education             0
      Self_Employed        32
      ApplicantIncome       0
      CoapplicantIncome     0
      LoanAmount           22
      Loan_Amount_Term     14
      Credit_History       50
      Property_Area         0
      Loan_Status           0
      dtype: int64
```

```
[25]: len(df)
```

```
[25]: 614
```

```
[26]: no_missing_values/len(df)
```

```
[26]: Loan_ID              0.000000
      Dependents           0.024430
      Education            0.000000
      Self_Employed        0.052117
      ApplicantIncome      0.000000
      CoapplicantIncome    0.000000
      LoanAmount           0.035831
      Loan_Amount_Term     0.022801
      Credit_History       0.081433
      Property_Area        0.000000
      Loan_Status          0.000000
      dtype: float64
```

```
[27]: no_missing_values/len(df)*100
```

```
[27]: Loan_ID              0.000000
      Dependents           2.442997
      Education            0.000000
      Self_Employed        5.211726
      ApplicantIncome      0.000000
      CoapplicantIncome    0.000000
      LoanAmount           3.583062
```

```
Loan_Amount_Term     2.280130
Credit_History       8.143322
Property_Area        0.000000
Loan_Status          0.000000
dtype: float64
```

[28]:

```
df.isnull().mean().round(4) * 100
```

[28]:

```
Loan_ID              0.00
Dependents           2.44
Education            0.00
Self_Employed        5.21
ApplicantIncome      0.00
CoapplicantIncome    0.00
LoanAmount           3.58
Loan_Amount_Term     2.28
Credit_History       8.14
Property_Area        0.00
Loan_Status          0.00
dtype: float64
```

[29]:

```
#https://towardsdatascience.com/
 ↪data-cleaning-in-python-the-ultimate-guide-2020-c63b88bf0a0d
```

[30]:

```
#https://medium.com/dunder-data/
 ↪finding-the-percentage-of-missing-values-in-a-pandas-dataframe-a04fa00f84ab
```