ex-1-data-cleaning

August 19, 2023

[1]:
```
import pandas as pd
import numpy as np
import seaborn as sns
```

[2]:
```
df = pd.read_csv("Loan_data.csv")
```

[3]:
```
import pandas as pd
df=pd.read_csv("Data_set.csv")
print(df)
df.head(10)
df.info()
df.isnull()
df.isnull().sum()
df['show_name']=df['show_name'].fillna(df['aired_on'].mode()[0])
df['aired_on']=df['aired_on'].fillna(df['aired_on'].mode()[0])
df['original_network']=df['original_network'].fillna(df['aired_on'].mode()[0])
df.head()

df['rating']=df['rating'].fillna(df['rating'].mean())
df['current_overall_rank']=df['current_overall_rank'].
 ↪fillna(df['current_overall_rank'].mean())
df.head()

df['watchers']=df['watchers'].fillna(df['watchers'].median())
df.head()

df.info()
df.isnull()
df.isnull().sum()
```

```
                   show_name      country  num_episodes             aired_on  \
0                        NaN  South Korea            16     Friday, Saturday
1                        NaN  South Korea            16     Friday, Saturday
2     Descendants of the Sun  South Korea            16  Wednesday, Thursday
3          Boys Over Flowers  South Korea            25      Monday, Tuesday
4                          W  South Korea            16  Wednesday, Thursday
..                       …          …           …                   …
95  Shut Up: Flower Boy Band  South Korea            16      Monday, Tuesday
```

```
96                    Blood  South Korea            20      Monday, Tuesday
97        Chicago Typewriter  South Korea            16     Friday, Saturday
98      Sungkyunkwan Scandal  South Korea            20      Monday, Tuesday
99                 Vagabond  South Korea            16     Friday, Saturday

   original_network  rating  current_overall_rank  lifetime_popularity_rank  \
0               tvN     8.9                  33.0                         1
1              jTBC     8.7                  89.0                         2
2              KBS2     8.7                  77.0                         3
3              KBS2     7.7                2249.0                         4
4               MBC     8.5                 201.0                         5
..              …     …                    …                          …
95              tvN     8.1                 806.0                        99
96             KBS2     7.4                3271.0                       100
97              tvN     8.8                  51.0                       101
98             KBS2     8.2                 605.0                       102
99     SBS, Netflix     8.5                 238.0                       103

    watchers
0   111706.0
1   100950.0
2    96318.0
3    94228.0
4    92121.0
..       …
95   34668.0
96   34666.0
97       NaN
98   34615.0
99   34523.0

[100 rows x 9 columns]
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100 entries, 0 to 99
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   show_name                 96 non-null     object
 1   country                   100 non-null    object
 2   num_episodes              100 non-null    int64
 3   aired_on                  99 non-null     object
 4   original_network          99 non-null     object
 5   rating                    96 non-null     float64
 6   current_overall_rank      97 non-null     float64
 7   lifetime_popularity_rank  100 non-null    int64
 8   watchers                  97 non-null     float64
dtypes: float64(3), int64(2), object(4)
memory usage: 7.2+ KB
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100 entries, 0 to 99
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   show_name                 100 non-null    object
 1   country                   100 non-null    object
 2   num_episodes              100 non-null    int64
 3   aired_on                  100 non-null    object
 4   original_network          100 non-null    object
 5   rating                    100 non-null    float64
 6   current_overall_rank      100 non-null    float64
 7   lifetime_popularity_rank  100 non-null    int64
 8   watchers                  100 non-null    float64
dtypes: float64(3), int64(2), object(4)
memory usage: 7.2+ KB
```

[3]:
```
show_name                   0
country                     0
num_episodes                0
aired_on                    0
original_network            0
rating                      0
current_overall_rank        0
lifetime_popularity_rank    0
watchers                    0
dtype: int64
```

[5]:
```
import pandas as pd
df=pd.read_csv("Data_set.csv")
print(df)
df.head(10)
```

```
                    show_name      country  num_episodes             aired_on  \
0                         NaN  South Korea            16     Friday, Saturday
1                         NaN  South Korea            16     Friday, Saturday
2     Descendants of the Sun   South Korea            16  Wednesday, Thursday
3          Boys Over Flowers   South Korea            25      Monday, Tuesday
4                           W  South Korea            16  Wednesday, Thursday
..                        …           …             …                   …
95  Shut Up: Flower Boy Band  South Korea            16      Monday, Tuesday
96                     Blood  South Korea            20      Monday, Tuesday
97        Chicago Typewriter  South Korea            16     Friday, Saturday
98      Sungkyunkwan Scandal  South Korea            20      Monday, Tuesday
99                  Vagabond  South Korea            16     Friday, Saturday

   original_network  rating  current_overall_rank  lifetime_popularity_rank  \
0               tvN     8.9                  33.0                         1
```

```
1              jTBC     8.7                89.0                        2
2              KBS2     8.7                77.0                        3
3              KBS2     7.7              2249.0                        4
4               MBC     8.5                201.0                       5
..              …      …                   …                         …
95              tvN     8.1                806.0                      99
96             KBS2     7.4              3271.0                      100
97              tvN     8.8                 51.0                     101
98             KBS2     8.2                605.0                     102
99     SBS, Netflix     8.5                238.0                     103

    watchers
0   111706.0
1   100950.0
2    96318.0
3    94228.0
4    92121.0
..       …
95   34668.0
96   34666.0
97       NaN
98   34615.0
99   34523.0

[100 rows x 9 columns]
```

[5]:

	show_name	country	num_episodes	aired_on	original_network	rating	current_overall_rank
0	NaN	South Korea	16	Friday, Saturday	tvN	8.9	33.0
1	NaN	South Korea	16	Friday, Saturday	jTBC	8.7	89.0
2	Descendants of the Sun	South Korea	16	Wednesday, Thursday	KBS2	8.7	77.0
3	Boys Over Flowers	South Korea	25	Monday, Tuesday	KBS2	7.7	2249.0
4	W	South Korea	16	Wednesday, Thursday	MBC	8.5	201.0
5	You Who Came from the Stars	South Korea	21	Wednesday, Thursday	SBS	8.6	112.0
6	Weightlifting Fairy Kim Bok Joo	South Korea	16	Wednesday, Thursday	MBC	8.8	40.0
7	The Heirs	South Korea	20	Wednesday, Thursday	SBS	7.5	2817.0
8	Pinocchio	South Korea	20				
9	Healer	South Korea	20				

```
8  Wednesday, Thursday              SBS     NaN                 273.0
9      Monday, Tuesday             KBS2     8.9                  25.0

   lifetime_popularity_rank  watchers
0                         1  111706.0
1                         2  100950.0
2                         3   96318.0
3                         4   94228.0
4                         5   92121.0
5                         6   91360.0
6                         7   91330.0
7                         8   90467.0
8                         9   82893.0
9                        10       NaN
```

[6]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100 entries, 0 to 99
Data columns (total 9 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   show_name                96 non-null     object
 1   country                  100 non-null    object
 2   num_episodes             100 non-null    int64
 3   aired_on                 99 non-null     object
 4   original_network         99 non-null     object
 5   rating                   96 non-null     float64
 6   current_overall_rank     97 non-null     float64
 7   lifetime_popularity_rank  100 non-null    int64
 8   watchers                 97 non-null     float64
dtypes: float64(3), int64(2), object(4)
memory usage: 7.2+ KB
```

[7]:
```
df.isnull()
```

[7]:
```
    show_name  country  num_episodes  aired_on  original_network  rating  \
0        True    False         False     False             False   False
1        True    False         False     False             False   False
2       False    False         False     False             False   False
3       False    False         False     False             False   False
4       False    False         False     False             False   False
..        …      …             …         …                 …       …
95      False    False         False     False             False   False
96      False    False         False     False             False   False
97      False    False         False     False             False   False
98      False    False         False     False             False   False
```

```
99      False    False         False     False             False   False

    current_overall_rank  lifetime_popularity_rank  watchers
0                  False                     False     False
1                  False                     False     False
2                  False                     False     False
3                  False                     False     False
4                  False                     False     False
..                   …                         …       …
95                 False                     False     False
96                 False                     False     False
97                 False                     False      True
98                 False                     False     False
99                 False                     False     False

[100 rows x 9 columns]
```

[8]:
```
df.isnull().sum()
```

[8]:
```
show_name                 4
country                   0
num_episodes              0
aired_on                  1
original_network          1
rating                    4
current_overall_rank      3
lifetime_popularity_rank  0
watchers                  3
dtype: int64
```

[9]:
```
df['show_name']=df['show_name'].fillna(df['aired_on'].mode()[0])
df['aired_on']=df['aired_on'].fillna(df['aired_on'].mode()[0])
df['original_network']=df['original_network'].fillna(df['aired_on'].mode()[0])
df.head()
```

[9]:
```
                show_name      country  num_episodes             aired_on  \
0     Wednesday, Thursday  South Korea            16     Friday, Saturday
1     Wednesday, Thursday  South Korea            16     Friday, Saturday
2  Descendants of the Sun  South Korea            16  Wednesday, Thursday
3       Boys Over Flowers  South Korea            25      Monday, Tuesday
4                       W  South Korea            16  Wednesday, Thursday

  original_network  rating  current_overall_rank  lifetime_popularity_rank  \
0              tvN     8.9                  33.0                         1
1             jTBC     8.7                  89.0                         2
2             KBS2     8.7                  77.0                         3
3             KBS2     7.7                2249.0                         4
```

```
4              MBC     8.5                  201.0                          5

   watchers
0  111706.0
1  100950.0
2   96318.0
3   94228.0
4   92121.0
```

[10]:
```
df['rating']=df['rating'].fillna(df['rating'].mean())
df['current_overall_rank']=df['current_overall_rank'].
 ↪fillna(df['current_overall_rank'].mean())
df.head()
```

[10]:
```
                show_name      country  num_episodes             aired_on  \
0     Wednesday, Thursday  South Korea            16     Friday, Saturday
1     Wednesday, Thursday  South Korea            16     Friday, Saturday
2  Descendants of the Sun  South Korea            16  Wednesday, Thursday
3       Boys Over Flowers  South Korea            25      Monday, Tuesday
4                       W  South Korea            16  Wednesday, Thursday

  original_network  rating  current_overall_rank  lifetime_popularity_rank  \
0              tvN     8.9                  33.0                         1
1             jTBC     8.7                  89.0                         2
2             KBS2     8.7                  77.0                         3
3             KBS2     7.7                2249.0                         4
4              MBC     8.5                 201.0                         5

   watchers
0  111706.0
1  100950.0
2   96318.0
3   94228.0
4   92121.0
```

[11]:
```
df['watchers']=df['watchers'].fillna(df['watchers'].median())
df.head()
```

[11]:
```
                show_name      country  num_episodes             aired_on  \
0     Wednesday, Thursday  South Korea            16     Friday, Saturday
1     Wednesday, Thursday  South Korea            16     Friday, Saturday
2  Descendants of the Sun  South Korea            16  Wednesday, Thursday
3       Boys Over Flowers  South Korea            25      Monday, Tuesday
4                       W  South Korea            16  Wednesday, Thursday

  original_network  rating  current_overall_rank  lifetime_popularity_rank  \
0              tvN     8.9                  33.0                         1
```

```
1             jTBC     8.7                89.0                       2
2             KBS2     8.7                77.0                       3
3             KBS2     7.7              2249.0                       4
4              MBC     8.5               201.0                       5

   watchers
0  111706.0
1  100950.0
2   96318.0
3   94228.0
4   92121.0
```

[12]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100 entries, 0 to 99
Data columns (total 9 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   show_name               100 non-null    object
 1   country                 100 non-null    object
 2   num_episodes            100 non-null    int64
 3   aired_on                100 non-null    object
 4   original_network        100 non-null    object
 5   rating                  100 non-null    float64
 6   current_overall_rank    100 non-null    float64
 7   lifetime_popularity_rank  100 non-null    int64
 8   watchers                100 non-null    float64
dtypes: float64(3), int64(2), object(4)
memory usage: 7.2+ KB
```

[13]:
```
df.isnull().sum()
```

[13]:
```
show_name                 0
country                   0
num_episodes              0
aired_on                  0
original_network          0
rating                    0
current_overall_rank      0
lifetime_popularity_rank  0
watchers                  0
dtype: int64
```

[14]:
```
import pandas as pd
df=pd.read_csv("Loan_data.csv")
print(df)
```

```
df.head(10)
```

```
      Loan_ID Gender Married Dependents     Education Self_Employed  \
0    LP001015   Male     Yes          0      Graduate            No
1    LP001022   Male     Yes          1      Graduate            No
2    LP001031   Male     Yes          2      Graduate            No
3    LP001035   Male     Yes          2      Graduate            No
4    LP001051   Male      No          0  Not Graduate            No
..        …    …     …        …           …            …
362  LP002971   Male     Yes         3+  Not Graduate           Yes
363  LP002975   Male     Yes          0      Graduate            No
364  LP002980   Male      No          0      Graduate            No
365  LP002986   Male     Yes          0      Graduate            No
366  LP002989   Male      No          0      Graduate           Yes

     ApplicantIncome  CoapplicantIncome  LoanAmount  Loan_Amount_Term  \
0               5720                  0       110.0             360.0
1               3076               1500       126.0             360.0
2               5000               1800       208.0             360.0
3               2340               2546       100.0             360.0
4               3276                  0        78.0             360.0
..               …                  …         …                …
362             4009               1777       113.0             360.0
363             4158                709       115.0             360.0
364             3250               1993       126.0             360.0
365             5000               2393       158.0             360.0
366             9200                  0        98.0             180.0

     Credit_History Property_Area
0               1.0         Urban
1               1.0         Urban
2               1.0         Urban
3               NaN         Urban
4               1.0         Urban
..              …           …
362             1.0         Urban
363             1.0         Urban
364             NaN     Semiurban
365             1.0         Rural
366             1.0         Rural

[367 rows x 12 columns]
```

[14]:
```
      Loan_ID  Gender Married Dependents     Education Self_Employed  \
 0  LP001015    Male     Yes          0      Graduate            No
 1  LP001022    Male     Yes          1      Graduate            No
 2  LP001031    Male     Yes          2      Graduate            No
```

```
3  LP001035    Male     Yes          2      Graduate            No
4  LP001051    Male      No          0  Not Graduate            No
5  LP001054    Male     Yes          0  Not Graduate           Yes
6  LP001055  Female      No          1  Not Graduate            No
7  LP001056    Male     Yes          2  Not Graduate            No
8  LP001059    Male     Yes          2      Graduate           NaN
9  LP001067    Male      No          0  Not Graduate            No

   ApplicantIncome  CoapplicantIncome  LoanAmount  Loan_Amount_Term  \
0             5720                  0       110.0             360.0
1             3076               1500       126.0             360.0
2             5000               1800       208.0             360.0
3             2340               2546       100.0             360.0
4             3276                  0        78.0             360.0
5             2165               3422       152.0             360.0
6             2226                  0        59.0             360.0
7             3881                  0       147.0             360.0
8            13633                  0       280.0             240.0
9             2400               2400       123.0             360.0

   Credit_History Property_Area
0             1.0         Urban
1             1.0         Urban
2             1.0         Urban
3             NaN         Urban
4             1.0         Urban
5             1.0         Urban
6             1.0     Semiurban
7             0.0         Rural
8             1.0         Urban
9             1.0     Semiurban
```

[15]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 367 entries, 0 to 366
Data columns (total 12 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Loan_ID            367 non-null    object
 1   Gender             356 non-null    object
 2   Married            367 non-null    object
 3   Dependents         357 non-null    object
 4   Education          367 non-null    object
 5   Self_Employed      344 non-null    object
 6   ApplicantIncome    367 non-null    int64
 7   CoapplicantIncome  367 non-null    int64
```

```
 8   LoanAmount          362 non-null    float64
 9   Loan_Amount_Term    361 non-null    float64
 10  Credit_History      338 non-null    float64
 11  Property_Area       367 non-null    object
dtypes: float64(3), int64(2), object(7)
memory usage: 34.5+ KB
```

[16]:
```
df['Loan_ID']=df['Loan_ID'].fillna(df['Dependents'].mode()[0])
df['Dependents']=df['Dependents'].fillna(df['Dependents'].mode()[0])
df['Education']=df['Education'].fillna(df['Dependents'].mode()[0])
df['Self_Employed']=df['Self_Employed'].fillna(df['Self_Employed'].mode()[0])
df['Gender']=df['Gender'].fillna(df['Gender'].mode()[0])
df.head()
```

[16]:

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area
0	LP001015	Male	Yes	0	Graduate	No	5720	0	110.0	360.0	1.0	Urban
1	LP001022	Male	Yes	1	Graduate	No	3076	1500	126.0	360.0	1.0	Urban
2	LP001031	Male	Yes	2	Graduate	No	5000	1800	208.0	360.0	1.0	Urban
3	LP001035	Male	Yes	2	Graduate	No	2340	2546	100.0	360.0	NaN	Urban
4	LP001051	Male	No	0	Not Graduate	No	3276	0	78.0	360.0	1.0	Urban

[17]:
```
df['ApplicantIncome']=df['ApplicantIncome'].fillna(df['ApplicantIncome'].mean())
df['Loan_Amount_Term']=df['Loan_Amount_Term'].fillna(df['Loan_Amount_Term'].
 ↪mean())
df['LoanAmount']=df['LoanAmount'].fillna(df['LoanAmount'].mean())
df.head()
```

[17]:

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed
0	LP001015	Male	Yes	0	Graduate	No
1	LP001022	Male	Yes	1	Graduate	No
2	LP001031	Male	Yes	2	Graduate	No
3	LP001035	Male	Yes	2	Graduate	No
4	LP001051	Male	No	0	Not Graduate	No

```
   ApplicantIncome  CoapplicantIncome  LoanAmount  Loan_Amount_Term  \
0             5720                  0       110.0             360.0
1             3076               1500       126.0             360.0
2             5000               1800       208.0             360.0
3             2340               2546       100.0             360.0
4             3276                  0        78.0             360.0

   Credit_History Property_Area
0             1.0         Urban
1             1.0         Urban
2             1.0         Urban
3             NaN         Urban
4             1.0         Urban
```

[18]:
```
df['Credit_History']=df['Credit_History'].fillna(df['Credit_History'].median())
df.head()
```

[18]:
```
    Loan_ID Gender Married Dependents     Education Self_Employed  \
0  LP001015   Male     Yes          0      Graduate            No
1  LP001022   Male     Yes          1      Graduate            No
2  LP001031   Male     Yes          2      Graduate            No
3  LP001035   Male     Yes          2      Graduate            No
4  LP001051   Male      No          0  Not Graduate            No

   ApplicantIncome  CoapplicantIncome  LoanAmount  Loan_Amount_Term  \
0             5720                  0       110.0             360.0
1             3076               1500       126.0             360.0
2             5000               1800       208.0             360.0
3             2340               2546       100.0             360.0
4             3276                  0        78.0             360.0

   Credit_History Property_Area
0             1.0         Urban
1             1.0         Urban
2             1.0         Urban
3             1.0         Urban
4             1.0         Urban
```

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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 367 entries, 0 to 366
Data columns (total 12 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Loan_ID            367 non-null    object
```

```
 1   Gender             367 non-null    object
 2   Married            367 non-null    object
 3   Dependents         367 non-null    object
 4   Education          367 non-null    object
 5   Self_Employed      367 non-null    object
 6   ApplicantIncome    367 non-null    int64
 7   CoapplicantIncome  367 non-null    int64
 8   LoanAmount         367 non-null    float64
 9   Loan_Amount_Term   367 non-null    float64
 10  Credit_History     367 non-null    float64
 11  Property_Area      367 non-null    object
dtypes: float64(3), int64(2), object(7)
memory usage: 34.5+ KB
```

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

[20]:
```
Loan_ID              0
Gender               0
Married              0
Dependents           0
Education            0
Self_Employed        0
ApplicantIncome      0
CoapplicantIncome    0
LoanAmount           0
Loan_Amount_Term     0
Credit_History       0
Property_Area        0
dtype: int64
```