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

96		Blood	South Korea	20	Monday, Tuesday
97	Chicago Ty	pewriter	South Korea	16	Friday, Saturday
98	Sungkyunkwan	Scandal	South Korea	20	Monday, Tuesday
99		Vagabond	South Korea	16	Friday, Saturday
	original_network	rating	<pre>current_overall_rank</pre>	lifetime	e_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

```
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
                                    100 non-null
         aired on
                                                     object
     4
         original_network
                                    100 non-null
                                                     object
     5
                                    100 non-null
                                                     float64
         rating
     6
                                    100 non-null
         current_overall_rank
                                                     float64
     7
         lifetime_popularity_rank
                                    100 non-null
                                                     int64
         watchers
                                    100 non-null
                                                     float64
    dtypes: float64(3), int64(2), object(4)
    memory usage: 7.2+ KB
[3]: show_name
                                  0
                                  0
     country
     num_episodes
                                  0
     aired_on
                                  0
     original_network
                                  0
                                  0
     rating
     current_overall_rank
                                  0
     lifetime_popularity_rank
                                  0
                                  0
     watchers
     dtype: int64
[5]: import pandas as pd
     df=pd.read_csv("Data_set.csv")
     print(df)
     df.head(10)
                                                 num_episodes
                                                                           aired_on \
                        show_name
                                        country
    0
                                   South Korea
                                                                   Friday, Saturday
                              {\tt NaN}
                                                           16
    1
                              {\tt NaN}
                                   South Korea
                                                           16
                                                                   Friday, Saturday
                                                               Wednesday, Thursday
    2
          Descendants of the Sun
                                   South Korea
                                                           16
                                                                    Monday, Tuesday
    3
               Boys Over Flowers
                                   South Korea
                                                           25
    4
                                   South Korea
                                                           16
                                                               Wednesday, Thursday
    95
        Shut Up: Flower Boy Band
                                   South Korea
                                                           16
                                                                    Monday, Tuesday
    96
                            Blood
                                   South Korea
                                                           20
                                                                    Monday, Tuesday
    97
                                   South Korea
                                                           16
                                                                   Friday, Saturday
              Chicago Typewriter
    98
            Sungkyunkwan Scandal
                                                           20
                                                                    Monday, Tuesday
                                   South Korea
    99
                         Vagabond
                                                                   Friday, Saturday
                                   South Korea
                                                           16
       original_network rating current_overall_rank lifetime_popularity_rank \
    0
                                                   33.0
                     tvN
                             8.9
                                                                                 1
```

<class 'pandas.core.frame.DataFrame'>

1 2 3 4	jTBC 8.7 KBS2 8.7 KBS2 7.7 MBC 8.5	89.0 77.0 2249.0 201.0)	2 3 4 5
95 96 97 98 99	tvN 8.1 KBS2 7.4 tvN 8.8 KBS2 8.2 SBS, Netflix 8.5	 806.0 3271.0 51.0 605.0 238.0)))	99 100 101 102 103
0 1 2 3 4 95 96 97 98 99	watchers 111706.0 100950.0 96318.0 94228.0 92121.0 34668.0 34666.0 NaN 34615.0 34523.0			
0 1 2	show_name NaN NaN Descendants of the Sun	country South Korea South Korea	num_episodes \ 16 16 16	
3 4	Boys Over Flowers W	South Korea South Korea	25 16	
5 6 7 8 9		South Korea South Korea South Korea South Korea South Korea	21 16 20 20 20	
0 1 2 3 4 5 6 7	aired_on original_net Friday, Saturday Friday, Saturday Wednesday, Thursday	twork rating tvN 8.9 jTBC 8.7 KBS2 7.7 KBS2 7.7 MBC 8.5 SBS 8.6 MBC 8.8 SBS 7.5	current_overall_rank	

[5]

	8	Wednesday, Thursday Monday, Tuesday	SB: KBS:				73.0 25.0	
	J	Honday, Tuebudy	KDD.	2 0.			20.0	
		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					
١.	дf	info()						
١.	uı.	11110()						
	<cl< td=""><td>ass 'pandas.core.frame.Dat</td><td>aFrame'></td><td></td><td></td><td></td><td></td><td></td></cl<>	ass 'pandas.core.frame.Dat	aFrame'>					
		geIndex: 100 entries, 0 to						
		a columns (total 9 columns						
	#	Column	Non-Null	Count	Dtype			
	0	show_name	96 non-n	ull	object			
	1	country	100 non-	null	object			
	2	num_episodes	100 non-	null	int64			
	3	aired_on	99 non-n	ull	object			
	4	original_network	99 non-n	ull	object			
	5	rating	96 non-n	ull	float64			
	6	current_overall_rank	97 non-n	ull	float64			
	7	lifetime_popularity_rank	100 non-	null	int64			
	8	watchers	97 non-n	ull	float64			
	dtyj	pes: float64(3), int64(2),	, object(4)					
	mem	ory usage: 7.2+ KB						
:	df.	isnull()						
:	_		-		_	al_network	_	\
	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	9	False	False	
		***				•••		
	95	False False	False	False		False	False	
	96	False False	False	False		False	False	
	97	False False	False	False	9	False	False	
	98	False False	False	False	Э	False	False	

[6]

[7]

[7]

	99	False	False	False	False	Fals	e False
	0 1 2 3 4 95 96 97 98	rrent_ove	False	lifetime_popu	Ilarity_rank False False False False False False False False	False	
[8]:	df.isn	ull().sum	()				
[8]:	rating curren	y isodes on al_networl t_overall me_popular rs	_rank	4 0 0 1 1 4 3 0 3			
[9]:	df['ai	red_on']=c iginal_ne	df['aired_c	n'].fillna(di	f['aired_on		on'].mode()[0])
[9]:	1 Desc 3 4 orig: 0 1	Wednesday cendants of Boys Ove inal_netwo	TBC 8.7	South Korea South Korea South Korea South Korea current_ove	erall_rank 33.0 89.0	16 Friday16 Friday16 Wednesday25 Monda	1 2
	2		3S2 8.7 3S2 7.7		77.0 2249.0		3 4

```
4
                     MBC
                                                                                  5
                              8.5
                                                   201.0
         watchers
      0 111706.0
        100950.0
      1
          96318.0
      2
      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]:
                                                                          aired on \
                      show name
                                      country num episodes
            Wednesday, Thursday
                                                                 Friday, Saturday
                                  South Korea
      1
            Wednesday, Thursday
                                  South Korea
                                                          16
                                                                 Friday, Saturday
         Descendants of the Sun
                                                              Wednesday, Thursday
                                  South Korea
                                                          16
              Boys Over Flowers
                                  South Korea
                                                          25
                                                                  Monday, Tuesday
      3
      4
                                  South Korea
                                                              Wednesday, Thursday
                                                          16
                                   current_overall_rank
                                                         lifetime_popularity_rank
        original_network
                          rating
      0
                              8.9
                                                    33.0
                     {\tt tvN}
                                                                                  2
                              8.7
                                                    89.0
      1
                    iTBC
      2
                    KBS2
                              8.7
                                                    77.0
                                                                                  3
      3
                    KBS2
                              7.7
                                                  2249.0
                                                                                  4
                     MBC
                              8.5
                                                   201.0
                                                                                  5
         watchers
        111706.0
      0
        100950.0
      1
      2
          96318.0
      3
          94228.0
          92121.0
[11]: df['watchers']=df['watchers'].fillna(df['watchers'].median())
      df.head()
Γ11]:
                       show name
                                      country num episodes
                                                                          aired on \
                                                                 Friday, Saturday
      0
            Wednesday, Thursday
                                  South Korea
                                                          16
      1
            Wednesday, Thursday
                                  South Korea
                                                          16
                                                                 Friday, Saturday
         Descendants of the Sun
                                  South Korea
                                                          16
                                                              Wednesday, Thursday
      3
              Boys Over Flowers
                                  South Korea
                                                          25
                                                                   Monday, Tuesday
      4
                                  South Korea
                                                          16
                                                              Wednesday, Thursday
                                  current_overall_rank lifetime_popularity_rank
        original_network rating
                              8.9
                                                    33.0
                     tvN
                                                                                  1
```

```
1
                    jTBC
                                                                                  2
      2
                              8.7
                                                   77.0
                                                                                  3
                    KBS2
      3
                                                                                  4
                    KBS2
                              7.7
                                                 2249.0
      4
                     MBC
                                                                                  5
                              8.5
                                                  201.0
         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
          show_name
      0
                                     100 non-null
                                                      object
      1
          country
                                     100 non-null
                                                      object
          num_episodes
      2
                                                      int64
                                     100 non-null
                                     100 non-null
          aired_on
                                                      object
          original_network
                                     100 non-null
                                                      object
      5
                                     100 non-null
                                                      float64
          rating
          current_overall_rank
      6
                                     100 non-null
                                                      float64
      7
          lifetime_popularity_rank 100 non-null
                                                      int64
                                     100 non-null
                                                      float64
          watchers
     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)
```

89.0

8.7

df.head(10)

```
Loan_ID Gender Married Dependents
                                                 Education Self_Employed
0
     LP001015
                 Male
                           Yes
                                          0
                                                  Graduate
                                                                        No
     LP001022
                                          1
1
                 Male
                           Yes
                                                  Graduate
                                                                        No
2
                                          2
     LP001031
                 Male
                           Yes
                                                  Graduate
                                                                        No
3
                                          2
     LP001035
                 Male
                           Yes
                                                  Graduate
                                                                        No
4
     LP001051
                 Male
                            No
                                             Not Graduate
                                                                        No
. .
                                             Not Graduate
362
     LP002971
                                         3+
                                                                       Yes
                 Male
                           Yes
363
     LP002975
                 Male
                           Yes
                                          0
                                                  Graduate
                                                                        No
     LP002980
                                          0
364
                 Male
                            No
                                                  {\tt Graduate}
                                                                        No
365
     LP002986
                 Male
                           Yes
                                          0
                                                  Graduate
                                                                        No
     LP002989
                                          0
366
                 Male
                            No
                                                  Graduate
                                                                       Yes
     ApplicantIncome
                        {\tt 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
                                                    78.0
                                                                       360.0
                                          0
. .
                  •••
                                                                       360.0
362
                 4009
                                       1777
                                                   113.0
363
                 4158
                                        709
                                                   115.0
                                                                       360.0
364
                 3250
                                       1993
                                                   126.0
                                                                       360.0
365
                 5000
                                       2393
                                                   158.0
                                                                       360.0
                 9200
                                          0
                                                    98.0
366
                                                                       180.0
     Credit_History Property_Area
                 1.0
0
                               Urban
1
                 1.0
                               Urban
2
                 1.0
                               Urban
3
                               Urban
                 NaN
4
                 1.0
                               Urban
. .
                 •••
362
                 1.0
                               Urban
363
                 1.0
                               Urban
                          Semiurban
364
                 NaN
                               Rural
365
                 1.0
366
                 1.0
                               Rural
[367 rows x 12 columns]
```

-

[14]:		${ t Loan_ID}$	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		Gradua	te No	
4	LP001051	Male	No	0	Not	Gradua	te No	
5	LP001054	Male	Yes	0	Not	Gradua	te Yes	
6	LP001055	Female	No	1	Not	Gradua	te No	
7	LP001056	Male	Yes	2	Not	Gradua	te No	
8	LP001059	Male	Yes	2		Gradua	te NaN	
9	LP001067	Male	No	0	Not	Gradua	te No	
	Applicant	Income	Coapplicant	Income	Loan.	Amount	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	1
9		2400		2400		123.0	360.0	1
	Credit_Hi	story Pi	roperty_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):

	***************************************	00111110,	
#	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
          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 \
      0 LP001015
                    Male
                             Yes
                                           0
                                                  Graduate
                                                                       No
      1 LP001022
                    Male
                             Yes
                                           1
                                                  Graduate
                                                                       No
                                           2
      2 LP001031
                    Male
                             Yes
                                                  Graduate
                                                                       No
                    Male
                             Yes
                                           2
      3 LP001035
                                                  Graduate
                                                                       No
      4 LP001051
                    Male
                              No
                                              Not Graduate
                                                                       No
                          CoapplicantIncome
                                              LoanAmount Loan_Amount_Term \
         ApplicantIncome
      0
                                                                      360.0
                    5720
                                                   110.0
      1
                    3076
                                        1500
                                                   126.0
                                                                      360.0
      2
                    5000
                                        1800
                                                   208.0
                                                                      360.0
      3
                                        2546
                                                   100.0
                                                                      360.0
                    2340
      4
                    3276
                                           0
                                                    78.0
                                                                      360.0
         Credit_History Property_Area
      0
                    1.0
                                Urban
                    1.0
      1
                                Urban
      2
                    1.0
                                Urban
      3
                    NaN
                                Urban
      4
                    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
      2 LP001031
                    Male
                             Yes
                                                  Graduate
                                                                      No
      3 LP001035
                    Male
                             Yes
                                           2
                                                  Graduate
                                                                       No
      4 LP001051
                    Male
                              No
                                           0 Not Graduate
                                                                       No
```

```
360.0
      0
                     5720
                                            0
                                                     110.0
                     3076
                                         1500
                                                     126.0
                                                                        360.0
      1
      2
                     5000
                                         1800
                                                     208.0
                                                                        360.0
                     2340
                                         2546
                                                     100.0
                                                                        360.0
      3
      4
                     3276
                                            0
                                                      78.0
                                                                        360.0
         Credit_History Property_Area
      0
                     1.0
                                  Urban
                     1.0
                                  Urban
      1
      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
                     Male
      2 LP001031
                              Yes
                                            2
                                                    Graduate
                                                                         No
      3 LP001035
                     Male
                              Yes
                                            2
                                                    Graduate
                                                                         No
      4 LP001051
                     Male
                               No
                                               Not Graduate
                                                                         No
         ApplicantIncome
                           CoapplicantIncome
                                               LoanAmount
                                                           Loan_Amount_Term \
                                                                        360.0
      0
                     5720
                                                     110.0
                     3076
      1
                                         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
                     1.0
                                  Urban
      1
      2
                     1.0
                                  Urban
      3
                     1.0
                                  Urban
                     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
          _____
          Loan ID
                               367 non-null
                                                object
```

CoapplicantIncome LoanAmount Loan_Amount_Term \

ApplicantIncome

```
Gender
                        367 non-null
                                        object
 1
 2
    Married
                        367 non-null
                                        object
 3
    Dependents
                        367 non-null
                                        object
 4
    Education
                        367 non-null
                                        object
 5
     Self_Employed
                                        object
                        367 non-null
 6
     ApplicantIncome
                        367 non-null
                                        int64
 7
     CoapplicantIncome
                        367 non-null
                                        int64
 8
    LoanAmount
                        367 non-null
                                        float64
    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()

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