Shahd Mohamed Amer

22010468 Ai

```
import pandas as pd
df=pd.read_csv(r'/content/books.csv')
df
```

	book_id	goodreads_book_id	best_book_id	work_id	books_count	isbn	isbn13	authors	original_publication_year	original_title	...	ratings_count	work_ratings_count	work_text_reviews_count	ratings_1	ratings_2	ratings_3	rat
0	1	2767052	2767052	2792775	272	439023483	9.780439e+12	Suzanne Collins	2008.0	The Hunger Games	...	4780653	4942365	155254	66715	127936	560092	1
1	2	3	3	4640799	491	439554934	9.780440e+12	J.K. Rowling, Mary GrandPré	1997.0	Harry Potter and the Philosopher's Stone	...	4602479	4800065	75867	75504	101676	455024	1
2	3	41865	41865	3212258	226	316015849	9.780316e+12	Stephenie Meyer	2005.0	Twilight	...	3866839	3916824	95009	456191	436802	793319	
3	6	11870085	11870085	16827462	226	525478817	9.780525e+12	John Green	2012.0	The Fault in Our Stars	...	2346404	2478609	140739	47994	92723	327550	
4	12	13335037	13335037	13155899	210	62024035	9.780062e+12	Veronica Roth	2011.0	Divergent	...	1903563	2216814	101023	36315	82870	310297	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1349	9925	86737	86737	3877568	52	1582349177	9.781582e+12	Mary Hoffman	2002.0	City of Masks	...	12048	13385	555	314	758	3154	
1350	9937	13010211	13010211	18171867	22	1595435712	9.781596e+12	Caragh M. O'Brien	2012.0	Promised	...	11766	12884	1260	256	1098	3565	
1351	9942	16074758	16074758	21869436	18	1442486557	9.781442e+12	Abigail Haas, Abby McDonald	2013.0	Dangerous Girls	...	10439	12970	2631	203	553	2029	
1352	9947	21393526	21393526	40690062	19	62320521	9.780062e+12	Maria Dahvana Headley	2015.0	Magonia	...	12510	13652	2910	577	1440	3881	
1353	9965	13065327	13065327	18230950	25	802734375	9.780803e+12	Simone Elkeles	2013.0	Wild Cards	...	13954	15400	1514	406	965	3331	

1354 rows × 23 columns

```
[ ] df.describe()
```

	book_id	goodreads_book_id	best_book_id	work_id	books_count	isbn13	original_publication_year	average_rating	ratings_count	work_ratings_count	work_text_reviews_count	ratings_1	ratings_2	ratings_3	ratings_
count	1354.000000	1.354000e+03	1.354000e+03	1.354000e+03	1354.000000	1.310000e+03	1351.000000	1354.000000	1.354000e+03	1.354000e+03	1354.000000	1354.000000	1354.000000	1354.000000	1.354000e+0
mean	4453.584195	5.951852e+06	6.120589e+06	8.707028e+06	50.330871	9.766700e+12	2003.422650	3.999357	9.160429e+04	9.915569e+04	5151.093058	2297.409158	5005.615953	17528.918021	3.060591e+0
std	2894.277455	6.664595e+06	6.935008e+06	9.813696e+06	61.338867	3.572069e+11	16.779301	0.224263	2.671266e+05	3.023637e+05	10730.335273	13708.587239	16259.838433	43549.306920	8.427851e+0
min	1.000000	1.000000e+00	1.000000e+00	1.150000e+02	1.000000	7.678361e+10	1868.000000	3.230000	6.221000e+03	8.833000e+03	49.000000	33.000000	133.000000	826.000000	1.660000e+0
25%	1850.250000	1.537868e+05	1.537962e+05	1.375035e+06	22.000000	9.780152e+12	2003.000000	3.850000	1.759325e+04	1.918150e+04	1162.500000	305.000000	978.000000	4140.500000	6.360500e+0
50%	4177.500000	3.305318e+06	3.422646e+06	4.005716e+06	37.000000	9.780440e+12	2006.000000	4.000000	2.943000e+04	3.255150e+04	2208.000000	619.000000	1732.500000	6557.000000	1.079550e+0
75%	6814.500000	9.917380e+06	1.019388e+07	1.435717e+07	58.000000	9.780805e+12	2011.000000	4.160000	6.073800e+04	6.681275e+04	4650.750000	1355.000000	3644.500000	13312.250000	2.227500e+0
max	9955.000000	3.207567e+07	3.360215e+07	4.963819e+07	1314.000000	9.788424e+12	2017.000000	4.740000	4.780653e+06	4.942365e+06	155254.000000	456191.000000	436802.000000	793319.000000	1.481305e+0

```
[ ] df.shape
```

```
(1354, 23)
```

```
[ ] columns_list = list(df.columns)
    columns_list
```

```
['book_id',
 'goodreads_book_id',
 'best_book_id',
 'work_id',
 'books_count',
 'isbn',
 'isbn13',
 'authors',
 'original_publication_year',
 'original_title',
 'title',
 'language_code',
 'average_rating',
 'ratings_count',
 'work_ratings_count',
 'work_text_reviews_count',
 'ratings_1',
```

```
[ ] df.duplicated().sum()
```

```
0
```

```
[ ] # check missing values in categorical variables
    df[categorical].isnull().sum()
```

```
isbn               52
authors             0
original_title     52
title               0
language_code     109
image_url           0
small_image_url     0
dtype: int64
```

```
[ ] # view frequency counts of values in categorical variables
    for var in categorical:
        print(df[var].value_counts())
```

```
isbn
439023483     1
689868235     1
1416914234    1
1423116186    1
670012092     1
             ..
451222385     1
60575808      1
1477823832    1
316127256     1
802734375     1
Name: count, Length: 1302, dtype: int64
authors
Meg Cabot                27
Tamora Pierce            25
L.J. Smith               15
John Flanagan            14
Rick Riordan             14
                         ..
Nicola Yoon               1
Barry Lyga                1
Annette Curtis Klause     1
Patricia Reilly Giff      1
Maria Dahvana Headley     1
Name: count, Length: 555, dtype: int64
original_title
Once                      2
Destined                  2
Twisted                   2
```

```
;0575808      1
.477823832    1
!16127256     1
;02734375     1
lame: count, Length: 1302, dtype: int64
uthors
leg Cabot                  27
'amora Pierce              25
..J. Smith                 15
lohn Flanagan              14
lick Riordan               14
                           ..
licola Yoon                 1
larry Lyga                  1
nnette Curtis Klause        1
'atricia Reilly Giff        1
laria Dahvana Headley       1
lame: count, Length: 555, dtype: int64
riginal_title
)nce                          2
)estined                      2
'wisted                       2
leaven                        2
)ark Reunion                  2
                             ..
'he Wrath & the Dawn          1
lild Magic (Immortals, #1)    1
'hysik                        1
 Hat Full of Sky              1
lild Cards                    1
lame: count, Length: 1289, dtype: int64
:itle
'he Hunger Games (The Hunger Games, #1)            1
)umplin' (Dumplin', #1)                            1
Jnspoken (The Lynburn Legacy, #1)                  1
'he Queen (The Selection, #0.4)                    1
:hasing Vermeer (Chasing Vermeer, #1)              1
                                                  ..
:ncarceron (Incarceron, #1)                        1
'he Awakening (The Vampire Diaries, #1)            1
:irelight (Firelight, #1)                          1
'he Face on the Milk Carton (Janie Johnson, #1)    1
lild Cards (Wild Cards, #1)                        1
lame: count, Length: 1354, dtype: int64
.anguage_code
:ng       852
:n-US     365
:n-GB      14
:n-CA      11
```

```
[ ] # view frequency distribution of categorical variables
    for var in categorical:
        frequency_distribution = df[var].value_counts() / float(len(df))
        print(f"Frequency distribution for {var}:")
        print(frequency_distribution, "\n")
```

```
Frequency distribution for isbn:
isbn
439023483     0.000739
689868235     0.000739
1416914234    0.000739
1423116186    0.000739
670012092     0.000739
                ...
451222385     0.000739
60575808      0.000739
1477823832    0.000739
316127256     0.000739
802734375     0.000739
Name: count, Length: 1302, dtype: float64

Frequency distribution for authors:
authors
Meg Cabot                 0.019941
Tamora Pierce             0.018464
L.J. Smith                0.011078
John Flanagan             0.010340
Rick Riordan              0.010340
                              ...
Nicola Yoon               0.000739
Barry Lyga                0.000739
Annette Curtis Klause     0.000739
Patricia Reilly Giff      0.000739
Maria Dahvana Headley     0.000739
Name: count, Length: 555, dtype: float64

Frequency distribution for original_title:
original_title
Once                              0.001477
Destined                          0.001477
Twisted                           0.001477
Heaven                            0.001477
Dark Reunion                      0.001477
                                      ...
The Wrath & the Dawn              0.000739
Wild Magic (Immortals, #1)        0.000739
Physik                            0.000739
A Hat Full of Sky                 0.000739
Wild Cards                        0.000739
Name: count, Length: 1289, dtype: float64
```

```
Once                                0.001477
Destined                            0.001477
Twisted                             0.001477
Heaven                              0.001477
Dark Reunion                        0.001477
                                      ...
The Wrath & the Dawn                0.000739
Wild Magic (Immortals, #1)          0.000739
Physik                              0.000739
A Hat Full of Sky                   0.000739
Wild Cards                          0.000739
Name: count, Length: 1289, dtype: float64

Frequency distribution for title:
title
The Hunger Games (The Hunger Games, #1)             0.000739
Dumplin' (Dumplin', #1)                             0.000739
Unspoken (The Lynburn Legacy, #1)                   0.000739
The Queen (The Selection, #0.4)                     0.000739
Chasing Vermeer (Chasing Vermeer, #1)               0.000739
                                                      ...
Incarceron (Incarceron, #1)                         0.000739
The Awakening (The Vampire Diaries, #1)             0.000739
Firelight (Firelight, #1)                           0.000739
The Face on the Milk Carton (Janie Johnson, #1)     0.000739
Wild Cards (Wild Cards, #1)                         0.000739
```

```
# check for cardinality in categorical variables
for var in categorical:
    print(var, ' contains ', len(df[var].unique()), ' labels')
```

```
isbn  contains  1303  labels
authors  contains  555  labels
original_title  contains  1290  labels
title  contains  1354  labels
language_code  contains  8  labels
image_url  contains  1014  labels
small_image_url  contains  1014  labels
```

```
# check missing values in categorical variables
df[categorical].isnull().sum()
```

```
isbn               52
authors             0
original_title     52
title               0
language_code     109
image_url           0
small_image_url     0
dtype: int64
```

```
[ ] # check missing values in numerical variables
    df[numerical].isnull().sum()
```

```
book_id                       0
goodreads_book_id             0
best_book_id                  0
work_id                       0
books_count                   0
isbn13                       44
original_publication_year     3
average_rating                0
ratings_count                 0
work_ratings_count            0
work_text_reviews_count       0
ratings_1                     0
ratings_2                     0
ratings_3                     0
ratings_4                     0
ratings_5                     0
dtype: int64
```

```
[ ] # Calculate the percentage of missing values in the numerical variables
    missing_values = df[numerical].isnull().sum()
    percentage_missing = ((missing_values / len(df)) * 100).sort_values(ascending=False)[missing_values>0]
    percentage_missing
```

```
isbn13                      3.249631
original_publication_year   0.221566
dtype: float64
```

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

```
book_id                       0
goodreads_book_id             0
best_book_id                  0
work_id                       0
books_count                   0
isbn                         52
isbn13                       44
authors                       0
original_publication_year     3
original_title               52
title                         0
language_code               109
average_rating                0
ratings_count                 0
work_ratings_count            0
work_text_reviews_count       0
ratings_1                     0
```

harrypotter

1	2	3	3	4640799	491	439554934	9.780440e+12	J.K. Rowling, Mary GrandPre	1997.0	Harry Potter and the Philosopher's Stone	...	4602479	4800065	75867	75504	101676	455024	1156318	3011543	assets.com/bo
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	J.K. Rowling, Mary GrandPré, Rufus Beck	[illegible]	Harry Potter and the Prisoner of Azkaban		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	J.K. Rowling, Mary GrandPré	[illegible]	Harry Potter and the Order of the Phoenix	...	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	assets.com/bo
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	Harry Potter and the Chamber of Secrets		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	J.K. Rowling, Mary GrandPre	[illegible]	Harry Potter and the Sorcerer's Stone		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	assets.com/bo
10	24	6	6	3046572	332	439139600	9.780439e+12	J.K. Rowling, Mary GrandPre	2000.0	Harry Potter and the Goblet of Fire	...	1753043	1868642	31084	6676	20210	151785	494926	1195045	assets.com/bo
11	25	136251	136251	2963218	263	545010225	9.780545e+12	J.K. Rowling, Mary GrandPre	2007.0	Harry Potter and the Deathly Hallows	...	1746574	1847395	51942	8353	22245	113646	383914	1318227	assets.com/bo
12	27	1	1	41335427	275	439785960	9.780440e+12	J.K. Rowling, Mary GrandPre	2005.0	Harry Potter and the Half-Blood Prince	...	1678823	1785676	27520	7308	21516	136333	459028	1161491	assets.com/bo
86	422	862041	862041	2962492	76	545044281	9.780545e+12	J.K. Rowling	1998.0	Complete Harry Potter Boxed Set	...	190050	204125	6508	1105	1285	7020	30666	164040	assets.com/bo
633	1755	10	10	21457570	6	439827604	9.780440e+12	J.K. Rowling	2005.0	Harry Potter Collection (Harry Potter, #1-6)		24618	26274	882	203	186	946	4891	21048	assets.com/bo
1030	7018	485440	485440	471792	42	0425198914	9.780425e+12	David Colbert	2001.0	The Magical Worlds of Harry Potter: A Treasury...	...	13620	15145	267	329	1125	3706	3595	6332	assets.com/asse

✓ Connected to Python 3 Google Compute Engine backend

```
columns = ['books_count','average_rating','ratings_count','work_ratings_count','work_text_reviews_count']
columnsdrop = df.columns.difference(columns)

# Drop all columns except the specified ones
harrypotter = harrypotter.drop(columns=columnsdrop)
harrypotter
```

	books_count	average_rating	ratings_count	work_ratings_count	work_text_reviews_count
1	491	4.44	4602479	4800065	75867
6	376	4.53	1832823	1969375	36099
8	307	4.46	1735368	1840548	28685
9	398	4.37	1779331	1906199	34172
10	332	4.53	1753043	1868642	31084
11	283	4.61	1746574	1847396	51942
12	275	4.54	1678823	1785676	27520
96	78	4.74	190050	204125	6508
613	8	4.73	24618	26274	882
1036	42	3.96	13820	15145	267
1266	5	4.40	10738	11732	185

```
the_most_selling_books = Harry_Potter.sort_values(by='work_ratings_count', ascending=False)
the_most_selling_books
```

	books_count	average_rating	ratings_count	work_ratings_count	work_text_reviews_count
1	491	4.44	4602479	4800065	75867
6	376	4.53	1832823	1969375	36099
9	398	4.37	1779331	1906199	34172
10	332	4.53	1753043	1868642	31084
11	283	4.61	1746574	1847396	51942
8	307	4.46	1735368	1840548	28685
12	275	4.54	1678823	1785676	27520
96	78	4.74	190050	204125	6508
613	8	4.73	24618	26274	882
1036	42	3.96	13820	15145	267
1266	5	4.40	10738	11732	185

```
avg_rate=the_most_selling_books['average_rating'].mean()
avg_rate
```

```
4.482727272727273
```

Code + Text

```
# Calculate the percentage of missing values in the categorical variables
missing_values = df[categorical].isnull().sum()
percentage_missing = ((missing_values / len(df)) * 100).sort_values(ascending=False)[missing_values>0]
percentage_missing
```

```
language_code     8.058222
isbn              3.848473
original_title    3.848473
dtype: float64
```

```
# find numerical variables
numerical = [var for var in df.columns if df[var].dtype!='O']
print('There are {} numerical variables\n'.format(len(numerical)))
print('The numerical variables are :', numerical)
```

```
There are 16 numerical variables

The numerical variables are : ['book_id', 'goodreads_book_id', 'best_book_id', 'work_id', 'books_count', 'isbn13', 'original_publication_year', 'average_rating', 'ratings_count', 'work_ratings_count', 'work_text_reviews_count', 'ratings_1', 'ratings_2', 'ratings_3',
```

```
# view the numerical variables
df[numerical].head()
```

	book_id	goodreads_book_id	best_book_id	work_id	books_count	isbn13	original_publication_year	average_rating	ratings_count	work_ratings_count	work_text_reviews_count	ratings_1	ratings_2	ratings_3	ratings_4	ratings_5
0	1	2767052	2767052	2792775	272	9.780439e+12	2008.0	4.34	4780653	4942365	155254	66715	127936	560092	1481305	2706317
1	2	3	3	4640799	491	9.780440e+12	1997.0	4.44	4602479	4800065	75867	75504	101676	455024	1156318	3011543
2	3	41865	41865	3212258	226	9.780316e+12	2005.0	3.57	3866839	3916824	95009	456191	436802	793319	875073	1355439
3	6	11870085	11870085	16827462	226	9.780525e+12	2012.0	4.26	2346404	2478609	140739	47994	92723	327550	698471	1311871
4	12	13335037	13335037	13155899	210	9.780062e+12	2011.0	4.24	1903563	2216814	101023	36315	82870	310297	673028	1114304

```
[ ] # find categorical variables
    categorical = [var for var in df.columns if df[var].dtype=='O']
    print('There are {} categorical variables\n'.format(len(categorical)))
    print('The categorical variables are :\n\n', categorical)
```

```
There are 7 categorical variables

The categorical variables are :

 ['isbn', 'authors', 'original_title', 'title', 'language_code', 'image_url', 'small_image_url']
```

```
[ ] # view the categorical variables
    df[categorical].head()
```

	isbn	authors	original_title	title	language_code	image_url	small_image_url
0	439023483	Suzanne Collins	The Hunger Games	The Hunger Games (The Hunger Games, #1)	eng	https://images.gr-assets.com/books/1447303603m...	https://images.gr-assets.com/books/1447303603s...
1	439554934	J.K. Rowling, Mary GrandPré	Harry Potter and the Philosopher's Stone	Harry Potter and the Sorcerer's Stone (Harry P...	eng	https://images.gr-assets.com/books/1474154022m...	https://images.gr-assets.com/books/1474154022s...
2	316015849	Stephenie Meyer	Twilight	Twilight (Twilight, #1)	en-US	https://images.gr-assets.com/books/1361039443m...	https://images.gr-assets.com/books/1361039443s...
3	525478817	John Green	The Fault in Our Stars	The Fault in Our Stars	eng	https://images.gr-assets.com/books/1360206420m...	https://images.gr-assets.com/books/1360206420s...
4	62024035	Veronica Roth	Divergent	Divergent (Divergent, #1)	eng	https://images.gr-assets.com/books/1328559506m...	https://images.gr-assets.com/books/1328559506s...

```
[ ] # check missing values in categorical variables
    df[categorical].isnull().sum()
```

```
[ ] df.head()
```

	book_id	goodreads_book_id	best_book_id	work_id	books_count	isbn	isbn13	authors	original_publication_year	original_title	...	ratings_count	work_ratings_count	work_text_reviews_count	ratings_1	ratings_2	ratings_3	ratings
0	1	2767052	2767052	2792775	272	439023483	9.780439e+12	Suzanne Collins	2008.0	The Hunger Games	...	4780653	4942365	155254	66715	127936	560092	1481
1	2	3	3	4640799	491	439554934	9.780440e+12	J.K. Rowling, Mary GrandPré	1997.0	Harry Potter and the Philosopher's Stone	...	4602479	4800065	75867	75504	101676	455024	1156
2	3	41865	41865	3212258	226	316015849	9.780316e+12	Stephenie Meyer	2005.0	Twilight	...	3866839	3916824	95009	456191	436802	793319	875
3	6	11870085	11870085	16827462	226	525478817	9.780525e+12	John Green	2012.0	The Fault in Our Stars	...	2346404	2478609	140739	47994	92723	327550	698
4	12	13335037	13335037	13155899	210	62024035	9.780062e+12	Veronica Roth	2011.0	Divergent	...	1903563	2216814	101023	36315	82870	310297	673

5 rows × 23 columns

```
[ ] df.tail()
```

	book_id	goodreads_book_id	best_book_id	work_id	books_count	isbn	isbn13	authors	original_publication_year	original_title	...	ratings_count	work_ratings_count	work_text_reviews_count	ratings_1	ratings_2	ratings_3	ra
1349	9925	86737	86737	3877968	52	1582349177	9.781582e+12	Mary Hoffman	2002.0	City of Masks	...	12048	13386	555	314	758	3154	
1350	9937	13010211	13010211	18171867	22	1596435712	9.781596e+12	Caragh M. O'Brien	2012.0	Promised	...	11766	12884	1260	256	1098	3665	
1351	9942	16074758	16074758	21869436	18	1442486597	9.781442e+12	Abigail Haas, Abby McDonald	2013.0	Dangerous Girls	...	10439	12970	2631	203	553	2029	
1352	9947	21393526	21393526	40690062	19	62320521	9.780062e+12	Maria Dahvana Headley	2015.0	Magonia	...	12510	13652	2511	577	1440	3881	
1353	9955	13065327	13065327	18230950	25	802734375	9.780803e+12	Simone	2013.0	Wild Cards	...	13954	15400	1514	406	965	3331	