In [1]:
```
import pandas as pd
```

In [26]:
```
books = pd.read_csv("ecommerce_book_data_2023_03_18.csv")
books
```

Out[26]:

	UPC	Book Title	Genre	Price	Availability	Star Rating
0	abeafe151a587d3b	The Long Shadow of Small Ghosts: Murder and Me...	Crime	10.97	15	1
1	88c21fcd38e2486e	Dark Notes	Erotica	19.19	15	5
2	bb8245f52c7cce8f	Amid the Chaos	Cultural	36.58	15	1
3	3968e3fbf4695d7c	Equal Is Unfair: America's Misguided Fight Aga...	Politics	56.86	12	1
4	2b5054a4192e9b06	Why the Right Went Wrong: Conservatism--From G...	Politics	52.65	14	4
...	...	...	...	...	...	...
995	a94350ee74deaa07	Under the Tuscan Sun	Travel	37.33	7	3
996	1809259a5a5f1d8d	Vagabonding: An Uncommon Guide to the Art of L...	Travel	36.94	8	2
997	f9705c362f070608	See America: A Celebration of Our National Par...	Travel	48.87	14	3
998	ce60436f52c5ee68	Full Moon over Noah’s Ark: An Odyssey to Mount...	Travel	49.43	15	4
999	a22124811bfa8350	It's Only the Himalayas	Travel	45.17	19	2

1000 rows × 6 columns

In [27]:
```
books.shape # rows, columns
```

Out[27]:
```
(1000, 6)
```

In [28]:
```
books.count() # checking records
```

Out[28]:
```
UPC             1000
Book Title      1000
Genre           1000
Price           1000
Availability    1000
Star Rating     1000
dtype: int64
```

In [29]:
```
books.isnull().sum() # checking null values
```

Out[29]:
```
UPC             0
Book Title      0
Genre           0
Price           0
Availability    0
Star Rating     0
dtype: int64
```

In [30]:
```
books.describe() # checking
```

Out[30]:

	Price	Availability	Star Rating
count	1000.00000	1000.000000	1000.000000
mean	35.07035	8.585000	2.923000
std	14.44669	5.654622	1.434967
min	10.00000	1.000000	1.000000
25%	22.10750	3.000000	2.000000
50%	35.98000	7.000000	3.000000
75%	47.45750	14.000000	4.000000
max	59.99000	22.000000	5.000000

In [12]:
```
books[["Genre"]].value_counts().head(10)
# top 10 genres with the most books available
```

Out[12]:
```
Genre
Default           152
Nonfiction        110
Sequential Art     75
Add a comment      67
Fiction            65
Young Adult        54
Fantasy            48
Romance            35
Mystery            32
Food and Drink     30
dtype: int64
```

In [19]:
```
books[["Genre"]].value_counts().tail(10)
# top 10 genres with the least books available
```

Out[19]:
```
Genre
Parenting        1
Paranormal       1
Short Stories    1
Novels           1
Crime            1
Suspense         1
Cultural         1
Erotica          1
Adult Fiction    1
Academic         1
dtype: int64
```

In [16]:
```
genre_book_count_average = books[["Genre"]].value_counts().mean()
# getting average of all genre's book count
print("The average book count overall by genre is:", genre_book_count_average)
```

```
The average book count overall by genre is: 20.0
```

In [55]:
```
books["Genre"].value_counts() >= 15
# checking all genres with greater than or equal to 15 books
```

Out[55]:
```
Default               True
Nonfiction            True
Sequential Art        True
Add a comment         True
Fiction               True
Young Adult           True
Fantasy               True
Romance               True
Mystery               True
Food and Drink        True
Childrens             True
Historical Fiction    True
Classics              True
Poetry                True
History               True
Horror                True
Womens Fiction        True
Science Fiction       True
Science              False
Music                False
Business             False
Travel               False
Philosophy           False
Thriller             False
Humor                False
Autobiography        False
Art                  False
Religion             False
Psychology           False
New Adult            False
Christian Fiction    False
Spirituality         False
Sports and Games     False
Biography            False
Self Help            False
Health               False
Politics             False
Contemporary         False
Christian            False
Historical           False
Paranormal           False
Short Stories        False
Cultural             False
Novels               False
Academic             False
Suspense             False
Erotica              False
Adult Fiction        False
Parenting            False
Crime                False
Name: Genre, dtype: bool
```

In [89]:
```
books[["Genre"]].value_counts().head(18)
# all genres with atleast 15 books
```

Out[89]:
```
Genre
Default                152
Nonfiction             110
Sequential Art          75
Add a comment           67
Fiction                 65
Young Adult             54
Fantasy                 48
Romance                 35
Mystery                 32
Food and Drink          30
Childrens               29
Historical Fiction      26
Poetry                  19
Classics                19
History                 18
Horror                  17
Womens Fiction          17
Science Fiction         16
dtype: int64
```

In [162]:
```
# create a dataframe with genres and their corresponding number of books
genre_data = {'genre': ["Default", "Nonfiction",
                        "Sequential Art", "Add a comment",
                        "Fiction", "Young Adult", "Fantasy",
                        "Romance", "Mystery", "Food and Drink",
                        "Childrens", "Historical Fiction", "Poetry",
                        "Classics", "History", "Horror",
                        "Womens Fiction", "Science Fiction"],
        'book_count': [152, 110,75,67,65,54,48,35,32,30,
                       29,26,19,19,18,17,17,16]}
genre_data_df = pd.DataFrame(genre_data)

random_books = pd.DataFrame()
# create an empty dataframe to hold the randomized data
```

In [163]:
```
# loop through the genres and randomly select books for each genre
for genre in genre_data_df["genre"]:
    records = genre_data_df.loc[genre_data_df['genre'] == genre,
                                'book_count'].values[0]
    samples_temp = books

    if records >= 20: # if genre has atleast 20 books
        samples_temp = books[books.Genre == genre].sample(n=20)
    elif records >= 15: # if a genre has 15-19 books
        samples_temp = books[books.Genre == genre].sample(n=records)

    random_books = pd.concat([random_books, samples_temp])
    # selected records are added to the dataframe

random_books.to_csv('randomized_books.csv', index=False)
# export records to .csv file
```

In [166]:
```
random_books[["Genre"]].value_counts()
# display the count of books per genres inside the dataframe
```

Out[166]:
```
Genre
Add a comment         20
Historical Fiction    20
Sequential Art        20
Romance               20
Nonfiction            20
Mystery               20
Childrens             20
Young Adult           20
Food and Drink        20
Fiction               20
Fantasy               20
Default               20
Poetry                19
Classics              19
History               18
Womens Fiction        17
Horror                17
Science Fiction       16
dtype: int64
```