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

In [2]:
```
data = pd.read_csv('books.csv')
```

In [48]:
```
data.head()
```

Out[48]:

	book_id	goodreads_book_id	best_book_id	work_id	books_count	isbn	isbn13	authors	original_publication_year	original_title	...
0	15	48855	48855	3532896	710	553296981	9.780553e+12	Anne Frank, Eleanor Roosevelt, B.M. Mooyaart-D...	1947.0	Het Achterhuis: Dagboekbrieven 14 juni 1942 - ...	...
1	40	19501	19501	3352398	185	143038419	9.780143e+12	Elizabeth Gilbert	2006.0	Eat, pray, love: one woman's search for everyt...	...
2	81	7445	7445	2944133	92	074324754X	9.780743e+12	Jeannette Walls	2005.0	The Glass Castle	...
3	82	1845	1845	3284484	108	385486804	9.780385e+12	Jon Krakauer	1996.0	Into the Wild	...
4	87	1617	1617	265616	109	374500010	9.780375e+12	Elie Wiesel, Marion Wiesel	1958.0	Un di Velt Hot Geshvign	...

5 rows × 25 columns

Q1: How many rows and columns are there in books.csv dataset?

In [4]:
```
data.shape
```

Out[4]: (399, 24)

There are 399 rows and 24 columns in the given dataset.

Q2: How many books do not have an original title?

In [5]:
```
data.isnull().sum()
```

Out[5]:
```
book_id                       0
goodreads_book_id             0
best_book_id                  0
work_id                       0
books_count                   0
isbn                         11
isbn13                       10
authors                       0
original_publication_year     0
original_title               36
title                         0
language_code                43
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
image_url                     0
small_image_url               0
NonEnglish                    0
dtype: int64
```

There are 36 books which do not an original title.

Q3: How many unique books are present in the dataset ? Evaluate based on the 'book_id' after removing null values in the original_title column.

In [13]:
```
# Deleting rows containing null values in the original_title column.

new_data = data[data['original_title'].notnull()]
```

In [15]:
```
new_data.head(4)
```

Out[15]:

	book_id	goodreads_book_id	best_book_id	work_id	books_count	isbn	isbn13	authors	original_publication_year	original_title	...
0	15	48855	48855	3532896	710	553296981	9.780553e+12	Anne Frank, Eleanor Roosevelt, B.M. Mooyaart-D...	1947.0	Het Achterhuis: Dagboekbrieven 14 juni 1942 - ...	...
1	40	19501	19501	3352398	185	143038419	9.780143e+12	Elizabeth Gilbert	2006.0	Eat, pray, love: one woman's search for everyt...	...
2	81	7445	7445	2944133	92	074324754X	9.780743e+12	Jeannette Walls	2005.0	The Glass Castle	...
3	82	1845	1845	3284484	108	385486804	9.780385e+12	Jon Krakauer	1996.0	Into the Wild	...

4 rows × 24 columns

In [17]:
```
new_data['book_id'].nunique()
```

Out[17]:
```
363
```

There are 363 unique books present in the dataset based on the 'book_id' after removing null values in the original_title column.

Q4: What is the average rating of all the books in the dataset based on ‘average_rating’?

In [23]:
```
data['average_rating'].mean()
```

Out[23]:
```
3.9837844611528843
```

The average rating of all the books in the dataset based on 'average_rating' is 3.98

Q5: Find the number of books published in the year ‘2000’ based on the ‘original_publication_year’.

In [25]:
```
required_data = data[data['original_publication_year']==2000]
```

In [32]:
```
required_data.head(4)
```

Out[32]:

	book_id	goodreads_book_id	best_book_id	work_id	books_count	isbn	isbn13	authors	original_publication_year	original_title	...
26	495	4953	4953	42857	65	375725784	9.780376e+12	Dave Eggers	2000.0	A Heartbreaking Work of Staggering Genius	...
28	574	33313	33313	4219	78	60899220	9.780061e+12	Anthony Bourdain	2000.0	Kitchen Confidential: Adventures in the Culina...	...
43	828	9516	9516	3303888	39	037571457X	9.780376e+12	Marjane Satrapi, Mattias Ripa	2000.0	Persepolis	...
117	2386	9522	9522	25686510	12	2844140580	9.782844e+12	Marjane Satrapi	2000.0	NaN	...

4 rows × 24 columns

In [33]:
```
required_data['book_id'].nunique()
```

Out[33]: 8

There are 8 books published in the year '2000' according to 'original_publication_year'.

Q6: Which book (title) has the maximum number of ratings based on 'work_ratings_count'.

In [35]:
```
data[data['work_ratings_count'] == data['work_ratings_count'].max()]
```

Out[35]:

	book_id	goodreads_book_id	best_book_id	work_id	books_count	isbn	isbn13	authors	original_publication_year	original_title	...
0	15	48855	48855	3532896	710	553296981	9.780553e+12	Anne Frank, Eleanor Roosevelt, B.M. Mooyaart-D...	1947.0	Het Achterhuis: Dagboekbrieven 14 juni 1942 - ...	...

1 rows × 24 columns

In [36]:
```
data[data['work_ratings_count'] == data['work_ratings_count'].max()]['title']
```

Out[36]:
```
0    The Diary of a Young Girl
Name: title, dtype: object
```

The book named 'The Diary of a Young Girl' has maximum number of ratings based on 'work_ratings_count'.

Q7: Bucket the average_rating of books into 11 buckets [0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0] with 0.5 decimal rounding (eg: average_rating 3.0 to 3.49 will fall in bucket 3.0). Plot bar graph to show total number of books in each rating bucket.

In [41]:
```
import matplotlib.pyplot as plt
import seaborn as sns


bins = [0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0]
labels = [f'{i}-{i+0.5}' for i in bins[:-1]]

 # Bucket the 'average_rating' column

data['rating_bucket'] = pd.cut(data['average_rating'], bins = bins, labels = labels, right = False)
```

In [42]:
```
data.head(4)
```

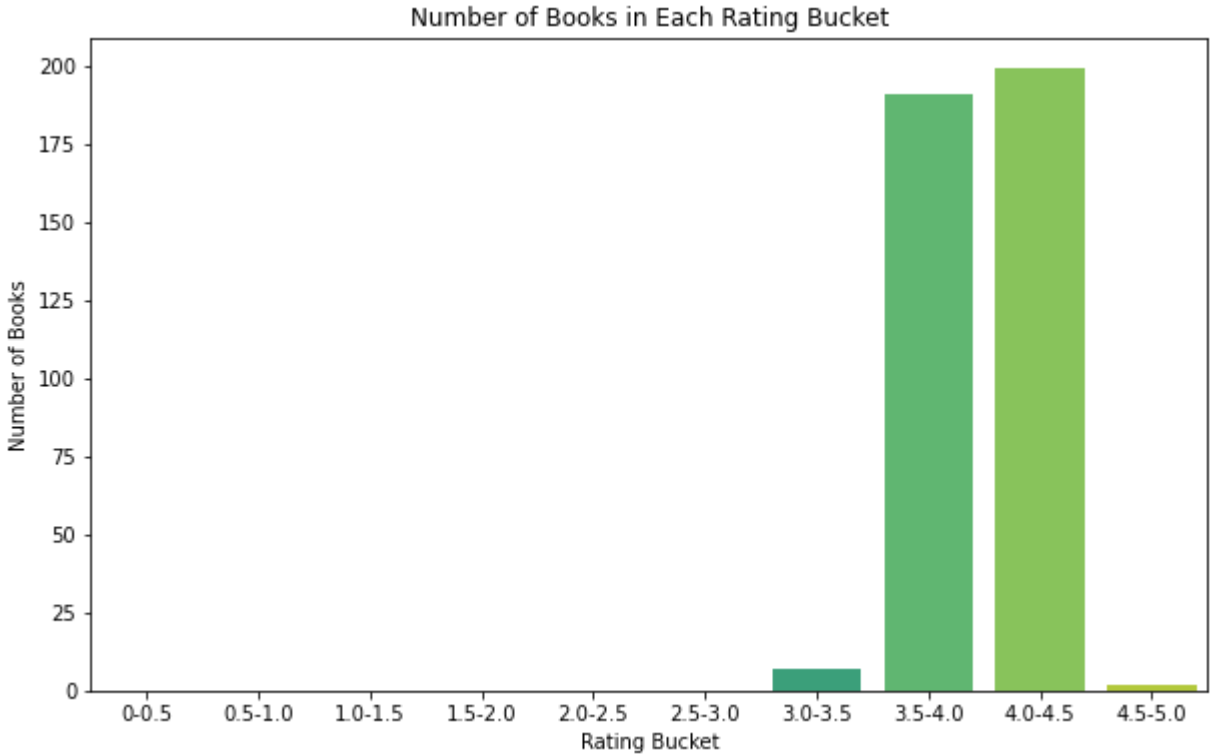
Out[42]:

	book_id	goodreads_book_id	best_book_id	work_id	books_count	isbn	isbn13	authors	original_publication_year	original_title	...
0	15	48855	48855	3532896	710	553296981	9.780553e+12	Anne Frank, Eleanor Roosevelt, B.M. Mooyaart-D...	1947.0	Het Achterhuis: Dagboekbrieven 14 juni 1942 - ...	...
1	40	19501	19501	3352398	185	143038419	9.780143e+12	Elizabeth Gilbert	2006.0	Eat, pray, love: one woman's search for everyt...	...
2	81	7445	7445	2944133	92	074324754X	9.780743e+12	Jeannette Walls	2005.0	The Glass Castle	...
3	82	1845	1845	3284484	108	385486804	9.780385e+12	Jon Krakauer	1996.0	Into the Wild	...

4 rows × 25 columns

In [47]:
```
# Plotting histogram

plt.figure(figsize=(10, 6))
sns.countplot(x = 'rating_bucket', data = data, palette = 'viridis')
plt.title('Number of Books in Each Rating Bucket')
plt.xlabel('Rating Bucket')
plt.ylabel('Number of Books')
plt.show()
```


Number of Books in Each Rating Bucket
Number of Books
200
175
150
125
100
75
50
25
0
0-0.5
0.5-1.0
1.0-1.5
1.5-2.0
2.0-2.5
2.5-3.0
3.0-3.5
3.5-4.0
4.0-4.5
4.5-5.0
Rating Bucket

In [ ]:

In [ ]: