In [75]:
```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
df_1=pd.read_csv("Customers.csv")
```

In [19]:
```
df_1.head(50)
```

Out[19]:

	CustomerID	CustomerName	Region	SignupDate
0	C0001	Lawrence Carroll	South America	2022-07-10
1	C0002	Elizabeth Lutz	Asia	2022-02-13
2	C0003	Michael Rivera	South America	2024-03-07
3	C0004	Kathleen Rodriguez	South America	2022-10-09
4	C0005	Laura Weber	Asia	2022-08-15
5	C0006	Brittany Palmer	South America	2024-01-07
6	C0007	Paul Graves	Asia	2022-06-18
7	C0008	David Li	North America	2024-01-13
8	C0009	Joy Clark	Europe	2023-08-14
9	C0010	Aaron Cox	Europe	2022-12-15
10	C0011	Bryan Mathews	South America	2022-12-12
11	C0012	Kevin May	South America	2024-08-07
12	C0013	Lauren Buchanan	South America	2024-05-19
13	C0014	Deborah Wilcox	Europe	2024-06-22
14	C0015	Tina Duran	North America	2023-11-20
15	C0016	Emily Woods	North America	2024-01-03
16	C0017	Jennifer King	Europe	2023-12-05
17	C0018	Tyler Haynes	North America	2024-09-21
18	C0019	Brandon Rodriguez	Europe	2023-01-12
19	C0020	Mr. Manuel Conway	North America	2024-06-11
20	C0021	Robert Blanchard	Asia	2023-04-17
21	C0022	Teresa Esparza	Asia	2023-10-27
22	C0023	Nicholas Cain	Europe	2022-03-04
23	C0024	Michele Cooley	North America	2024-02-05
24	C0025	Gregory Odom	South America	2022-07-04
25	C0026	Sara Miller	North America	2024-05-03
26	C0027	Justin Heath	Asia	2022-11-09
27	C0028	Jennifer Pena	Asia	2024-06-29
28	C0029	Erin Manning	North America	2022-04-16
29	C0030	Mark Brock	North America	2024-01-30
30	C0031	Tina Miller	South America	2024-04-11
31	C0032	Dustin Campbell	South America	2024-04-17
32	C0033	Tyler Holt	North America	2024-08-04
33	C0034	Dalton Perez	North America	2023-09-27
34	C0035	Brianna Richardson	North America	2024-10-01
35	C0036	Brian Aguilar DDS	North America	2024-07-06
36	C0037	Linda Smith	Europe	2023-02-04
37	C0038	Jeffrey Perkins	North America	2022-04-16
38	C0039	Angela Harris	South America	2024-10-13

	CustomerID	CustomerName	Region	SignupDate
39	C0040	Michael Harrell	Asia	2022-03-07
40	C0041	Lindsey Deleon	Europe	2023-12-27
41	C0042	Heather Riley	North America	2023-03-15
42	C0043	Sandy Short MD	Asia	2023-02-05
43	C0044	Kenneth Alexander	Europe	2024-07-10
44	C0045	Michael Williams	Asia	2022-02-25
45	C0046	Beth Cardenas	North America	2024-10-23
46	C0047	Samantha Frank	North America	2024-03-22
47	C0048	Matthew Park	South America	2024-11-07
48	C0049	Jason Yates	North America	2024-09-18
49	C0050	Ryan Davis	North America	2024-03-02

In [10]:
```
df_1.dtypes
```

Out[10]:
```
CustomerID      object
CustomerName    object
Region          object
SignupDate      object
dtype: object
```

In [11]:
```
df_1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 4 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   CustomerID    200 non-null    object
 1   CustomerName  200 non-null    object
 2   Region        200 non-null    object
 3   SignupDate    200 non-null    object
dtypes: object(4)
memory usage: 6.4+ KB
```

In [12]:
```
df_1.isnull().sum()
```

Out[12]:
```
CustomerID      0
CustomerName    0
Region          0
SignupDate      0
dtype: int64
```

In [13]:
```
df_1.nunique()
```

Out[13]:
```
CustomerID      200
CustomerName    200
Region            4
SignupDate      179
dtype: int64
```

In [16]:
```
df_1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 4 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   CustomerID    200 non-null    object
 1   CustomerName  200 non-null    object
 2   Region        200 non-null    object
 3   SignupDate    200 non-null    object
dtypes: object(4)
memory usage: 6.4+ KB
```

In [15]:
```
df_1.shape
```

Out[15]:
```
(200, 4)
```

In [20]:
```
"""
Customers.csv

No.of.Columns : 4
Column names : CustomerID,Customer Name,Region,signupdate
Regions(4) : Asia,North America,South America,Europe
signupdate range : 2022 - 2024 , few customers signed up on same date
Customer Id : Primary key


"""
```

Out[20]:
```
'\nCustomers.csv\n\nNo.of.Columns : 4\nColumn names : CustomerID,Customer
Name,Region,signupdate\nRegions(4) : Asia,North America,South America,Euro
pe\nsignupdate range : 2022 - 2024\nCustomer Id : Primary key\n\n'
```

In [21]:
```
df_2 = pd.read_csv("Products.csv")
```

In [31]:
```
df_2.head(10)
```

Out[31]:

	ProductID	ProductName	Category	Price
0	P001	ActiveWear Biography	Books	169.30
1	P002	ActiveWear Smartwatch	Electronics	346.30
2	P003	ComfortLiving Biography	Books	44.12
3	P004	BookWorld Rug	Home Decor	95.69
4	P005	TechPro T-Shirt	Clothing	429.31
5	P006	ActiveWear Rug	Home Decor	121.32
6	P007	SoundWave Cookbook	Books	420.15
7	P008	BookWorld Bluetooth Speaker	Electronics	146.85
8	P009	BookWorld Wall Art	Home Decor	325.01
9	P010	ComfortLiving Smartwatch	Electronics	350.13

In [22]:
```
df_2.isnull().sum()
```

Out[22]:
```
ProductID      0
ProductName    0
Category       0
Price          0
dtype: int64
```

In [27]:
```
df_2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100 entries, 0 to 99
Data columns (total 4 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   ProductID    100 non-null    object 
 1   ProductName  100 non-null    object 
 2   Category     100 non-null    object 
 3   Price        100 non-null    float64
dtypes: float64(1), object(3)
memory usage: 3.3+ KB
```

In [28]:
```
df_2.dtypes
```

Out[28]:
```
ProductID       object
ProductName     object
Category        object
Price          float64
dtype: object
```

In [48]:
```
df_2['ProductName'].value_counts()
```

Out[48]:
```
ProductName
ActiveWear Smartwatch         4
SoundWave Headphones          4
BookWorld Biography           3
TechPro T-Shirt               3
SoundWave Cookbook            3
                             ..
BookWorld Jacket              1
ComfortLiving Smartphone      1
SoundWave T-Shirt             1
ComfortLiving Cookware Set    1
SoundWave Laptop              1
Name: count, Length: 66, dtype: int64
```

In [56]:
```
filtered_rows = df_2[df_2['ProductName'] =="SoundWave Headphones"]
filtered_rows
```

Out[56]:

	ProductID	ProductName	Category	Price
26	P027	SoundWave Headphones	Electronics	229.06
30	P031	SoundWave Headphones	Electronics	196.40
87	P088	SoundWave Headphones	Electronics	263.55
95	P096	SoundWave Headphones	Electronics	307.47

In [30]:
```
df_2.nunique()
```

Out[30]:
```
ProductID      100
ProductName     66
Category         4
Price          100
dtype: int64
```

In [32]:
```
df_2['Price'].max()
```

Out[32]:
```
497.76
```

In [34]:
```
df_2['Price'].min()
```

Out[34]:
```
16.08
```

In [ ]:
```
"""
Products.csv

No.of.Columns : 4
Column names : ProductId,Product Name,Category,Price
Categoy(4) : Books,Electromics,Home decor,Clothing
Price range : 16-498
Product Id : Primary key
No .of.unique products : 66

"""
```

In [35]:
```
df_3=pd.read_csv("transactions.csv")
```

In [36]:
```
df_3.isnull().sum()
```

Out[36]:
```
TransactionID      0
CustomerID         0
ProductID          0
TransactionDate    0
Quantity           0
TotalValue         0
Price              0
dtype: int64
```

In [37]:
```
df_3.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 7 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   TransactionID    1000 non-null   object 
 1   CustomerID       1000 non-null   object 
 2   ProductID        1000 non-null   object 
 3   TransactionDate  1000 non-null   object 
 4   Quantity         1000 non-null   int64  
 5   TotalValue       1000 non-null   float64
 6   Price            1000 non-null   float64
dtypes: float64(2), int64(1), object(4)
memory usage: 54.8+ KB
```

In [38]:
```
df_3.head(20)
```

Out[38]:

	TransactionID	CustomerID	ProductID	TransactionDate	Quantity	TotalValue	Price
0	T00001	C0199	P067	2024-08-25 12:38:23	1	300.68	300.68
1	T00112	C0146	P067	2024-05-27 22:23:54	1	300.68	300.68
2	T00166	C0127	P067	2024-04-25 07:38:55	1	300.68	300.68
3	T00272	C0087	P067	2024-03-26 22:55:37	2	601.36	300.68
4	T00363	C0070	P067	2024-03-21 15:10:10	3	902.04	300.68
5	T00442	C0188	P067	2024-12-26 14:40:03	1	300.68	300.68
6	T00490	C0195	P067	2024-11-24 11:49:48	3	902.04	300.68
7	T00536	C0008	P067	2024-09-22 06:13:59	1	300.68	300.68
8	T00564	C0157	P067	2024-12-07 17:57:40	3	902.04	300.68
9	T00631	C0130	P067	2024-05-14 23:14:59	2	601.36	300.68
10	T00727	C0051	P067	2024-01-20 04:52:14	3	902.04	300.68
11	T00729	C0075	P067	2024-10-07 06:38:36	2	601.36	300.68
12	T00797	C0155	P067	2024-10-04 20:42:53	2	601.36	300.68
13	T00002	C0092	P034	2024-12-16 03:31:07	2	434.10	217.05
14	T00063	C0088	P034	2024-06-19 07:16:55	2	434.10	217.05
15	T00093	C0109	P034	2024-08-11 04:10:13	4	868.20	217.05
16	T00127	C0041	P034	2024-06-03 21:07:56	4	868.20	217.05
17	T00270	C0101	P034	2024-11-07 02:48:08	3	651.15	217.05
18	T00358	C0154	P034	2024-04-08 20:47:01	1	217.05	217.05
19	T00518	C0200	P034	2024-12-11 03:05:50	4	868.20	217.05

In [39]:
```
df_3.nunique()
```

Out[39]:
```
TransactionID      1000
CustomerID          199
ProductID           100
TransactionDate    1000
Quantity              4
TotalValue          369
Price               100
dtype: int64
```

In [61]:
```
"""
Transactions.csv

Total no.of.transactions : 1000
Column names : TransactionID,CustomerID,ProductID,TransactionDate,Quantity,
No.of customers invloved in 1000 transactions : 199
No.of customers invloved in 1000 transactions : 100
No.of quanties in invloved in 1000 transactions: 1-4
No .of.unique products : 66

"""
```

Out[61]:
```
'\nTransactions.csv\n\nTotal no.of.transactions : 1000\nColumn names : Tra
nsactionID,CustomerID,ProductID,TransactionDate,Quantity,TotalValue,Price
\nNo.of customers invloved in 1000 transactions : 199\nNo.of customers inv
loved in 1000 transactions : 100\nNo.of quanties in invloved in 1000 trans
actions: 1-4\nNo .of.unique products : 66\n\n'
```

In [62]:
```
#EXPLORATORY DATA ANALYSIS

#name of products that sold a lot
#category with more no.of.products
#range of price of categories
#customer name with more transactions
#product sold a lot
```

In [60]:
```
merged_data = df_3.merge(df_1, on='CustomerID').merge(df_2, on='ProductID')
merged_data
```

Out[60]:

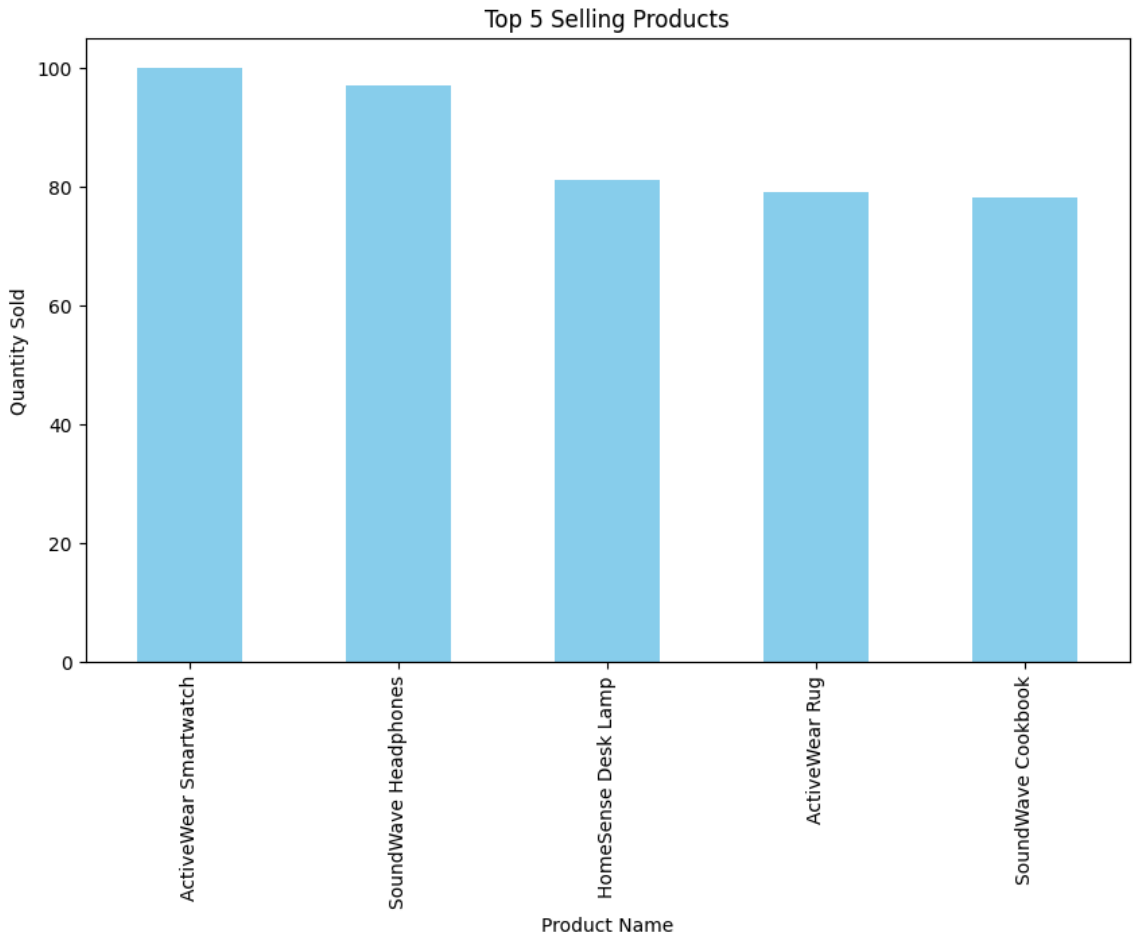
	TransactionID	CustomerID	ProductID	TransactionDate	Quantity	TotalValue	Price_x	C
0	T00001	C0199	P067	2024-08-25 12:38:23	1	300.68	300.68	/
1	T00112	C0146	P067	2024-05-27 22:23:54	1	300.68	300.68	I
2	T00166	C0127	P067	2024-04-25 07:38:55	1	300.68	300.68	K
3	T00272	C0087	P067	2024-03-26 22:55:37	2	601.36	300.68	T
4	T00363	C0070	P067	2024-03-21 15:10:10	3	902.04	300.68	
...	...	...	...	...	...	...	...	
995	T00630	C0031	P093	2024-10-08 23:58:14	2	609.88	304.94	
996	T00672	C0165	P044	2024-07-28 00:09:49	4	75.28	18.82	
997	T00711	C0165	P044	2024-06-11 15:51:14	4	75.28	18.82	
998	T00878	C0165	P044	2024-09-24 21:15:21	3	56.46	18.82	
999	T00157	C0169	P044	2024-11-09 09:07:36	2	37.64	18.82	

1000 rows × 13 columns

In [63]:
```
top_selling_products = merged_data.groupby('ProductName')['Quantity'].sum()
print("Top 5 Products Sold:")
print(top_selling_products)
```

```
Top 5 Products Sold:
ProductName
ActiveWear Smartwatch    100
SoundWave Headphones      97
HomeSense Desk Lamp       81
ActiveWear Rug            79
SoundWave Cookbook        78
Name: Quantity, dtype: int64
```

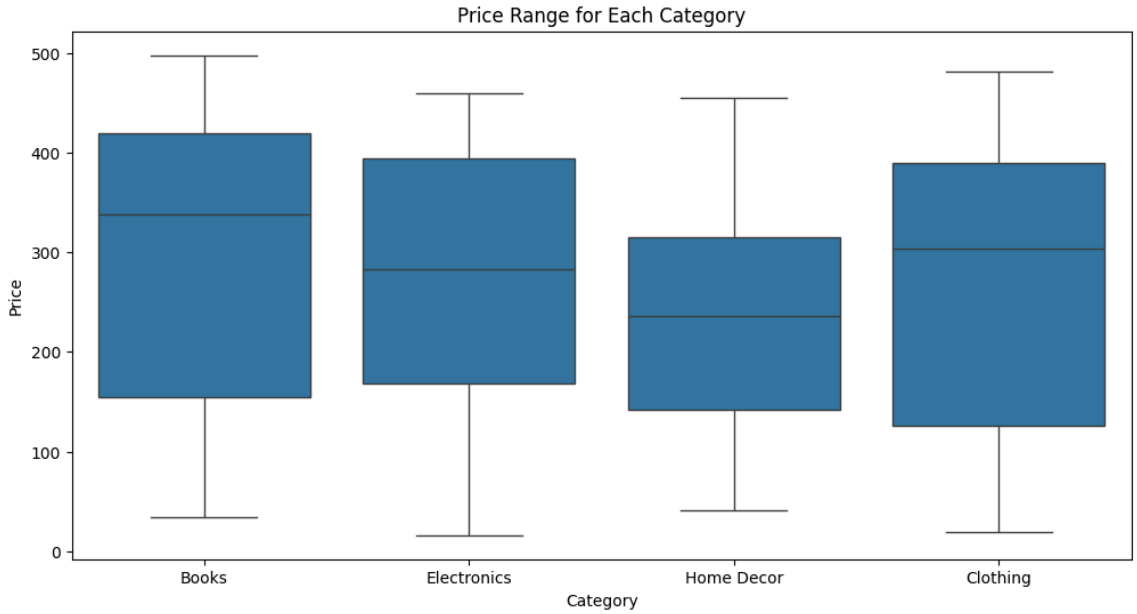
In [65]:
```
plt.figure(figsize=(10, 6))
top_selling_products.plot(kind='bar', color='skyblue')
plt.title('Top 5 Selling Products')
plt.xlabel('Product Name')
plt.ylabel('Quantity Sold')
plt.show()
```



In [72]:
```
category_product_count = df_2['Category'].value_counts().sort_values(ascend
print(category_product_count)
```

```
Category
Books          26
Electronics    26
Clothing       25
Home Decor     23
Name: count, dtype: int64
```

In [78]:
```
plt.figure(figsize=(12, 6))
sns.boxplot(x='Category', y='Price', data=df_2)
plt.title('Price Range for Each Category')
plt.xlabel('Category')
plt.ylabel('Price')
plt.show()
```



In [79]:
```
top_customer_transactions = merged_data.groupby('CustomerName')['Transactio
print("\nCustomer with the most transactions:")
print(top_customer_transactions)
```

```
Customer with the most transactions:
CustomerName
William Adams    11
Name: TransactionID, dtype: int64
```

In [80]:
```
most_sold_product = merged_data.groupby('ProductName')['Quantity'].sum().id
print("\nProduct that sold the most:", most_sold_product)
```

```
Product that sold the most: ActiveWear Smartwatch
```