Importing Libraries

In [1]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

Loading the Dataset

In [2]:
```
df = pd.read_csv("C:\\Users\\rakhi\\Downloads\\shopping Trends Analysis\\shopping_trends.csv")
df
```

Out[2]:

	Customer ID	Age	Gender	Item Purchased	Category	Purchase Amount (USD)	Location	Size	Color	Season	Review Rating	Subscription Status	Payment Method	S
0	1	55	Male	Blouse	Clothing	53	Kentucky	L	Gray	Winter	3.1	Yes	Credit Card	
1	2	19	Male	Sweater	Clothing	64	Maine	L	Maroon	Winter	3.1	Yes	Bank Transfer	
2	3	50	Male	Jeans	Clothing	73	Massachusetts	S	Maroon	Spring	3.1	Yes	Cash	
3	4	21	Male	Sandals	Footwear	90	Rhode Island	M	Maroon	Spring	3.5	Yes	PayPal	
4	5	45	Male	Blouse	Clothing	49	Oregon	M	Turquoise	Spring	2.7	Yes	Cash	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
3895	3896	40	Female	Hoodie	Clothing	28	Virginia	L	Turquoise	Summer	4.2	No	Cash	
3896	3897	52	Female	Backpack	Accessories	49	Iowa	L	White	Spring	4.5	No	PayPal	
3897	3898	46	Female	Belt	Accessories	33	New Jersey	L	Green	Spring	2.9	No	Credit Card	
3898	3899	44	Female	Shoes	Footwear	77	Minnesota	S	Brown	Summer	3.8	No	PayPal	
3899	3900	52	Female	Handbag	Accessories	81	California	M	Beige	Spring	3.1	No	Bank Transfer	

3900 rows × 19 columns

In [3]:
```
df.head()
```

Out[3]:

	Customer ID	Age	Gender	Item Purchased	Category	Purchase Amount (USD)	Location	Size	Color	Season	Review Rating	Subscription Status	Payment Method	Shippin Typ
0	1	55	Male	Blouse	Clothing	53	Kentucky	L	Gray	Winter	3.1	Yes	Credit Card	Expres
1	2	19	Male	Sweater	Clothing	64	Maine	L	Maroon	Winter	3.1	Yes	Bank Transfer	Expres
2	3	50	Male	Jeans	Clothing	73	Massachusetts	S	Maroon	Spring	3.1	Yes	Cash	Fre Shippir
3	4	21	Male	Sandals	Footwear	90	Rhode Island	M	Maroon	Spring	3.5	Yes	PayPal	Next Da A
4	5	45	Male	Blouse	Clothing	49	Oregon	M	Turquoise	Spring	2.7	Yes	Cash	Fre Shippir

In [4]:
```
df.tail()
```

Out[4]:

	Customer ID	Age	Gender	Item Purchased	Category	Purchase Amount (USD)	Location	Size	Color	Season	Review Rating	Subscription Status	Payment Method	Shipp T
3895	3896	40	Female	Hoodie	Clothing	28	Virginia	L	Turquoise	Summer	4.2	No	Cash	2-I Shipp
3896	3897	52	Female	Backpack	Accessories	49	Iowa	L	White	Spring	4.5	No	PayPal	S Pic
3897	3898	46	Female	Belt	Accessories	33	New Jersey	L	Green	Spring	2.9	No	Credit Card	Stanc
3898	3899	44	Female	Shoes	Footwear	77	Minnesota	S	Brown	Summer	3.8	No	PayPal	Expr
3899	3900	52	Female	Handbag	Accessories	81	California	M	Beige	Spring	3.1	No	Bank Transfer	S Pic

In [5]:
```
# Shape of the dataset
```

In [5]:
```
# Shape of the dataset
df.shape
```

Out[5]:
```
(3900, 19)
```

In [6]:
```
df.dtypes
```

Out[6]:
```
Customer ID                 int64
Age                         int64
Gender                     object
Item Purchased             object
Category                   object
Purchase Amount (USD)       int64
Location                   object
Size                       object
Color                      object
Season                     object
Review Rating             float64
Subscription Status        object
Payment Method             object
Shipping Type              object
Discount Applied           object
Promo Code Used            object
Previous Purchases          int64
Preferred Payment Method   object
Frequency of Purchases     object
dtype: object
```

In [7]:
```
# column in the dataset
df.columns
```

Out[7]:
```
Index(['Customer ID', 'Age', 'Gender', 'Item Purchased', 'Category',
       'Purchase Amount (USD)', 'Location', 'Size', 'Color', 'Season',
       'Review Rating', 'Subscription Status', 'Payment Method',
       'Shipping Type', 'Discount Applied', 'Promo Code Used',
       'Previous Purchases', 'Preferred Payment Method',
       'Frequency of Purchases'],
      dtype='object')
```

In [8]:
```
# Information about the dataset
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3900 entries, 0 to 3899
Data columns (total 19 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Customer ID               3900 non-null   int64
 1   Age                       3900 non-null   int64
 2   Gender                    3900 non-null   object
 3   Item Purchased            3900 non-null   object
 4   Category                  3900 non-null   object
 5   Purchase Amount (USD)     3900 non-null   int64
 6   Location                  3900 non-null   object
 7   Size                      3900 non-null   object
 8   Color                     3900 non-null   object
 9   Season                    3900 non-null   object
 10  Review Rating             3900 non-null   float64
 11  Subscription Status       3900 non-null   object
 12  Payment Method            3900 non-null   object
 13  Shipping Type             3900 non-null   object
 14  Discount Applied          3900 non-null   object
 15  Promo Code Used           3900 non-null   object
 16  Previous Purchases        3900 non-null   int64
 17  Preferred Payment Method  3900 non-null   object
 18  Frequency of Purchases    3900 non-null   object
dtypes: float64(1), int64(4), object(14)
memory usage: 579.0+ KB
```

In [9]:
```
df.describe()
```

Out[9]:

	Customer ID	Age	Purchase Amount (USD)	Review Rating	Previous Purchases
count	3900.000000	3900.000000	3900.000000	3900.000000	3900.000000
mean	1950.500000	44.068462	59.764359	3.749949	25.351538
std	1125.977353	15.207589	23.685392	0.716223	14.447125
min	1.000000	18.000000	20.000000	2.500000	1.000000
25%	975.750000	31.000000	39.000000	3.100000	13.000000
50%	1950.500000	44.000000	60.000000	3.700000	25.000000
75%	2925.250000	57.000000	81.000000	4.400000	38.000000
max	3900.000000	70.000000	100.000000	5.000000	50.000000

In [10]:
```
# checking if there are any null values present in the dataset or not ?
df.isnull().sum()
```

Out[10]:
```
Customer ID                 0
Age                         0
Gender                      0
Item Purchased              0
Category                    0
Purchase Amount (USD)       0
Location                    0
Size                        0
Color                       0
Season                      0
Review Rating               0
Subscription Status         0
Payment Method              0
Shipping Type               0
Discount Applied            0
Promo Code Used             0
Previous Purchases          0
Preferred Payment Method    0
Frequency of Purchases      0
dtype: int64
```

In [11]:
```
# checking if there are any duplicate values present in the dataset or not ?
df.duplicated().sum()
```

Out[11]:
```
0
```

In [12]:
```
# number of unique values of each column
df.nunique()
```

Out[12]:
```
Customer ID                 3900
Age                           53
Gender                         2
Item Purchased                25
Category                       4
Purchase Amount (USD)         81
Location                      50
Size                           4
Color                         25
Season                         4
Review Rating                 26
Subscription Status            2
Payment Method                 6
Shipping Type                  6
Discount Applied               2
Promo Code Used                2
Previous Purchases            50
Preferred Payment Method       6
Frequency of Purchases         7
dtype: int64
```

In [14]:
```
df["Gender"].value_counts()
```

Out[14]:
```
Male      2652
Female    1248
Name: Gender, dtype: int64
```

In [16]:
```
df["Category"].value_counts()
```

Out[16]:
```
Clothing       1737
Accessories    1240
Footwear        599
Outerwear       324
Name: Category, dtype: int64
```

In [18]:
```
df['Item Purchased'].value_counts()
```

Out[18]:
```
Blouse        171
Jewelry       171
Pants         171
Shirt         169
Dress         166
Sweater       164
Jacket        163
Belt          161
Sunglasses    161
Coat          161
Sandals       160
Socks         159
Skirt         158
Shorts        157
Scarf         157
Hat           154
Handbag       153
Hoodie        151
Shoes         150
T-shirt       147
Sneakers      145
Boots         144
Backpack      143
Gloves        140
Jeans         124
Name: Item Purchased, dtype: int64
```

In [19]:
```
# Average age of customers

average_age = df["Age"].mean()
print("Average Age:", average_age)
```

```
Average Age: 44.06846153846154
```

In [20]:
```
# Most common item purchased

most_common_item = df['Item Purchased'].mode()
print("Most Common Item Purchased:", most_common_item)
```

```
Most Common Item Purchased: 0     Blouse
1    Jewelry
2      Pants
Name: Item Purchased, dtype: object
```

In [21]:
```
# Total Purchase amount for each category

total_purchase_by_category = df.groupby('Category')['Purchase Amount (USD)'].sum()
print("Total Purchase Amount by Category:")
print(total_purchase_by_category)
```

```
Total Purchase Amount by Category:
Category
Accessories     74200
Clothing       104264
Footwear        36093
Outerwear       18524
Name: Purchase Amount (USD), dtype: int64
```

In [22]:
```
# Average review rating for male customers and female customers

average_rating_male = df[df['Gender'] == 'Male']['Review Rating'].mean()
average_rating_female = df[df['Gender'] == 'Female']['Review Rating'].mean()
print("Average Review Rating for Male Customers:",average_rating_male)
print("Average Review Rating dor Female Customers:",average_rating_female)
```

```
Average Review Rating for Male Customers: 3.7539592760180995
Average Review Rating dor Female Customers: 3.741426282051282
```

In [24]:
```
# Most Common Payment method used by customers

most_common_payment_method = df['Payment Method'].mode()[0]
print("Most Common Payment Method:", most_common_payment_method)
```

```
Most Common Payment Method: Credit Card
```

In [26]:
```
# Median purchase amount(USD)

median_purchase_amount = df['Purchase Amount (USD)'].median()
print("Median Purchase Amount (USD):", median_purchase_amount)
```

```
Median Purchase Amount (USD): 60.0
```

In [27]:
```
# How many customers have opted for the Subscription

subscription_count = df[df['Subscription Status'] == 'Yes']['Customer ID'].count()
print("Number of Customers with Scubscription: ",subscription_count)
```

```
Number of Customers with Scubscription:  1053
```

In [29]:
```
# Average purchase amount for customers with a subscription status of 'Yes' or 'No'
```

```
avg_purchase_subscription_yes = df[df['Subscription Status'] == 'Yes']['Purchase Amount (USD)'].mean()
avg_purchase_subscription_no = df[df['Subscription Status'] == 'No']['Purchase Amount (USD)'].mean()
print("Average Purchase Amount For Subscription 'Yes':",avg_purchase_subscription_yes)
print("Average Purchase Amount For Subscription 'No':",avg_purchase_subscription_no)
```

```
Average Purchase Amount For Subscription 'Yes': 59.49192782526116
Average Purchase Amount For Subscription 'No': 59.865121180189675
```

In [30]:
```
# Average age of males and females who purchased from each category

pd.crosstab(df['Gender'], df['Category'], values=df['Age'],aggfunc=np.average)
```

Out[30]:

Category	Accessories	Clothing	Footwear	Outerwear
Gender				
Female	44.283163	43.620504	44.482412	44.128713
Male	44.196934	43.859441	44.422500	44.394619

In [31]:
```
# Number of items purchased from each category

pd.crosstab(df['Category'],df['Item Purchased']).T
```

Out[31]:

Category	Accessories	Clothing	Footwear	Outerwear
Item Purchased				
Backpack	143	0	0	0
Belt	161	0	0	0
Blouse	0	171	0	0
Boots	0	0	144	0
Coat	0	0	0	161
Dress	0	166	0	0
Gloves	140	0	0	0
Handbag	153	0	0	0
Hat	154	0	0	0
Hoodie	0	151	0	0
Jacket	0	0	0	163
Jeans	0	124	0	0
Jewelry	171	0	0	0
Pants	0	171	0	0
Sandals	0	0	160	0
Scarf	157	0	0	0
Shirt	0	169	0	0
Shoes	0	0	150	0
Shorts	0	157	0	0
Skirt	0	158	0	0
Sneakers	0	0	145	0
Socks	0	159	0	0
Sunglasses	161	0	0	0
Sweater	0	164	0	0
T-shirt	0	147	0	0

In [32]:
```
# Maxium and Minimum review rating

max_review_rating = df['Review Rating'].max()
min_review_rating = df['Review Rating'].min()
print("Maximum Review Rating:",max_review_rating)
print("Minimum Review Rating:",min_review_rating)
```

```
Maximum Review Rating: 5.0
Minimum Review Rating: 2.5
```

In [33]:
```
category = df['Category'].unique()
category
```

Out[33]:
```
array(['Clothing', 'Footwear', 'Outerwear', 'Accessories'], dtype=object)
```

In [34]:
```
df['Color'].value_counts().nlargest(5)
```

Out[34]:
```
Olive     177
Yellow    174
Silver    173
Teal      172
Green     169
Name: Color, dtype: int64
```

In [35]:
```
df.groupby('Location')['Purchase Amount (USD)'].mean().sort_values(ascending = False)
```

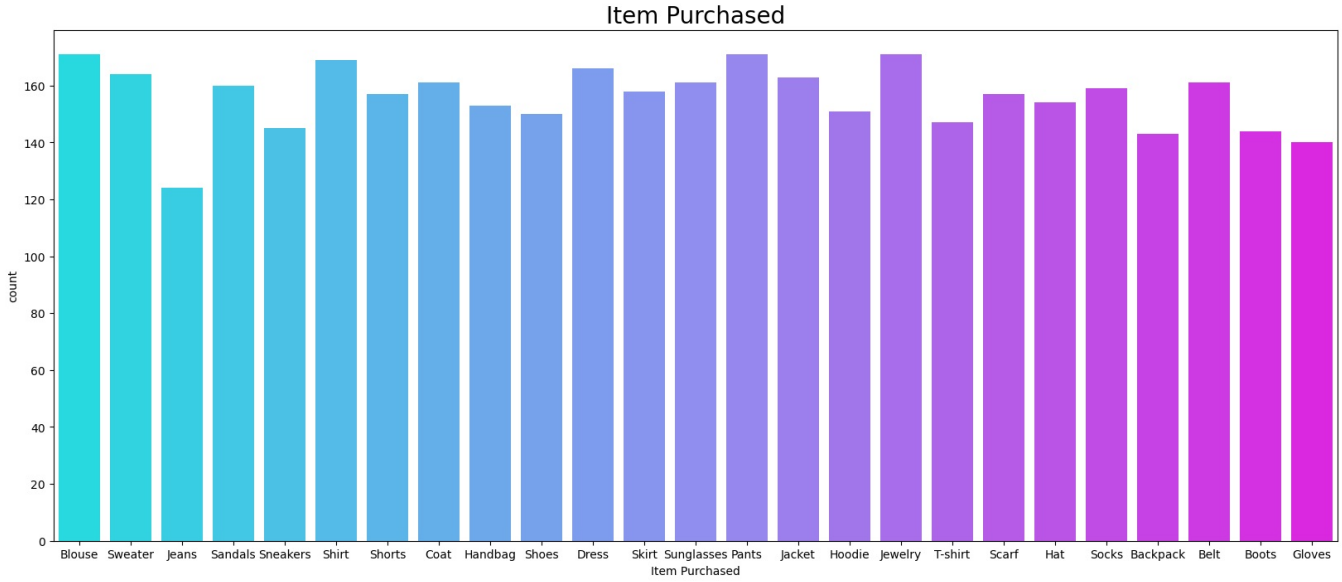
Out[35]:
```
Location
Alaska            67.597222
Pennsylvania      66.567568
Arizona           66.553846
West Virginia     63.876543
Nevada            63.379310
Washington        63.328767
North Dakota      62.891566
Virginia          62.883117
Utah              62.577465
Michigan          62.095890
Tennessee         61.974026
New Mexico        61.901235
Rhode Island      61.444444
Texas             61.194805
Arkansas          61.113924
Illinois          61.054348
Mississippi       61.037500
Massachusetts     60.888889
Iowa              60.884058
North Carolina    60.794872
Wyoming           60.690141
South Dakota      60.514286
New York          60.425287
Ohio              60.376623
Montana           60.250000
Idaho             60.075269
Nebraska          59.448276
New Hampshire     59.422535
Alabama           59.112360
California        59.000000
Indiana           58.924051
Georgia           58.797468
South Carolina    58.407895
Oklahoma          58.346667
Missouri          57.913580
Hawaii            57.723077
Louisiana         57.714286
Oregon            57.337838
Vermont           57.176471
Maine             56.987013
New Jersey        56.746269
Minnesota         56.556818
Colorado          56.293333
Wisconsin         55.946667
Florida           55.852941
Maryland          55.755814
Kentucky          55.721519
Delaware          55.325581
Kansas            54.555556
Connecticut       54.179487
Name: Purchase Amount (USD), dtype: float64
```

In [38]:
```
df.head()
```

Out[38]:

	Customer ID	Age	Gender	Item Purchased	Category	Purchase Amount (USD)	Location	Size	Color	Season	Review Rating	Subscription Status	Payment Method	Shippin Typ
0	1	55	Male	Blouse	Clothing	53	Kentucky	L	Gray	Winter	3.1	Yes	Credit Card	Expres
1	2	19	Male	Sweater	Clothing	64	Maine	L	Maroon	Winter	3.1	Yes	Bank Transfer	Expres
2	3	50	Male	Jeans	Clothing	73	Massachusetts	S	Maroon	Spring	3.1	Yes	Cash	Fre Shippir
3	4	21	Male	Sandals	Footwear	90	Rhode Island	M	Maroon	Spring	3.5	Yes	PayPal	Next Da A
4	5	45	Male	Blouse	Clothing	49	Oregon	M	Turquoise	Spring	2.7	Yes	Cash	Fre Shippir

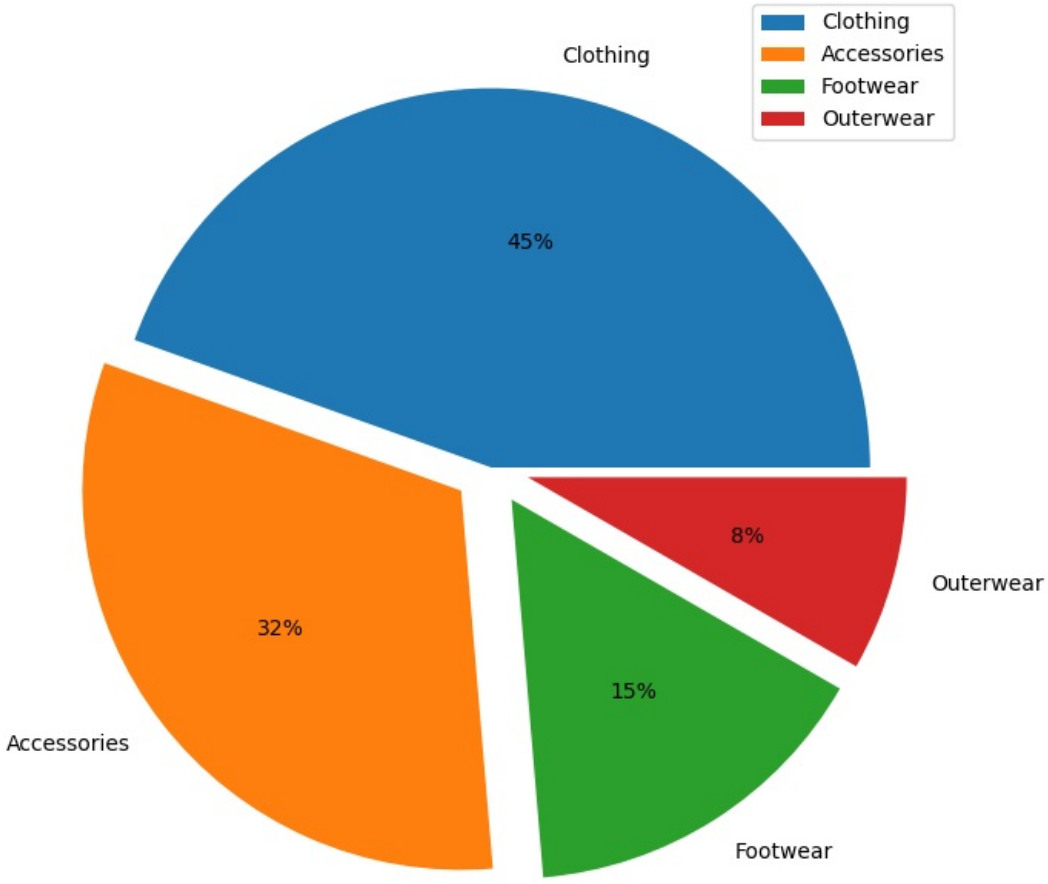
In [39]:
```
plt.subplots(figsize=(20,8), dpi=100)
sns.countplot(data= df, x='Item Purchased',palette='cool')
plt.title("Item Purchased",fontsize=20)
plt.show()
```



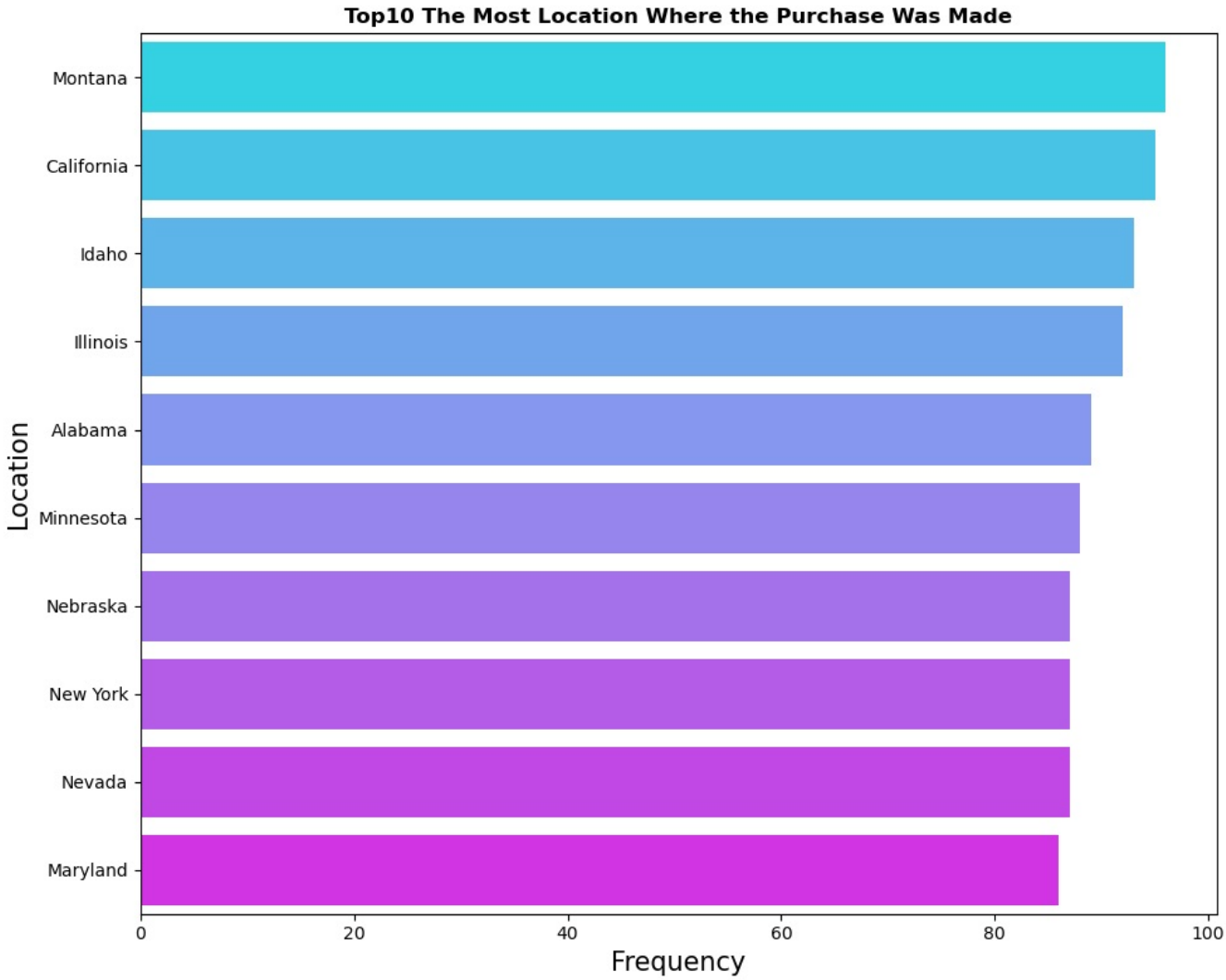
In [40]:
```
CountofCategory = df['Category'].value_counts()
CountofCategory
```

Out[40]:
```
Clothing       1737
Accessories    1240
Footwear        599
Outerwear       324
Name: Category, dtype: int64
```

In [41]:
```
plt.figure(figsize=(8,8))
plt.pie(CountofCategory,labels=CountofCategory.index,autopct='%0.0f%%',explode=(0,0.1,0.1,0.1))
plt.legend(CountofCategory.index,loc =1)
plt.show()
```



In [43]:
```
top_10_Location = df.Location.value_counts().sort_values(ascending=False)[:10]
plt.figure(figsize=(10,8))
sns.barplot(x=top_10_Location, y=top_10_Location.index,palette='cool',linewidth = 4)
plt.title('Top10 The Most Location Where the Purchase Was Made',loc='center',fontweight='bold',fontsize=12)
plt.xlabel('Frequency',fontsize=15)
plt.ylabel('Location',fontsize=15)
plt.tight_layout()
plt.show()
```

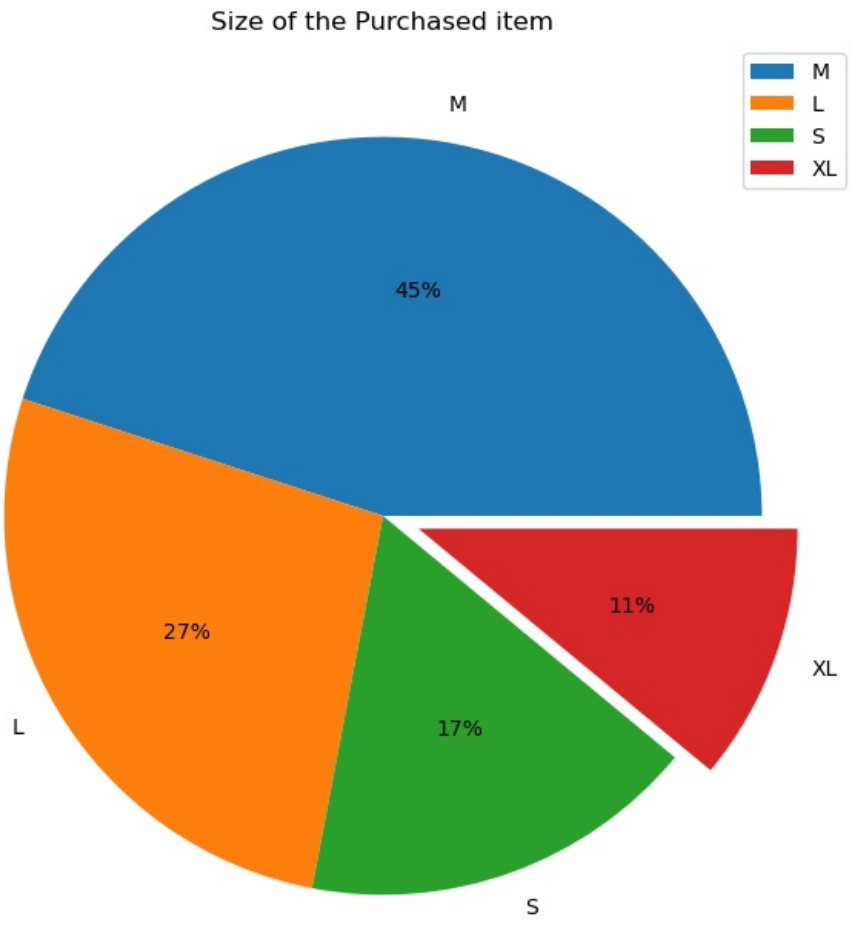


In [47]:
```
size_count = df['Size'].value_counts()
size_count
```

Out[47]:
```
M     1755
L     1053
S      663
XL     429
Name: Size, dtype: int64
```

In [53]:
```
import matplotlib as plt
import matplotlib.pyplot as plt
plt.figure(figsize=(8,8))
plt.pie(size_count,labels=['M','L','S','XL'],autopct='%0.0f%%',explode=(0,0,0,0.1))
plt.legend(['M','L','S','XL'],loc =1)
plt.title('Size of the Purchased item')
```

Out[53]:
```
Text(0.5, 1.0, 'Size of the Purchased item')
```



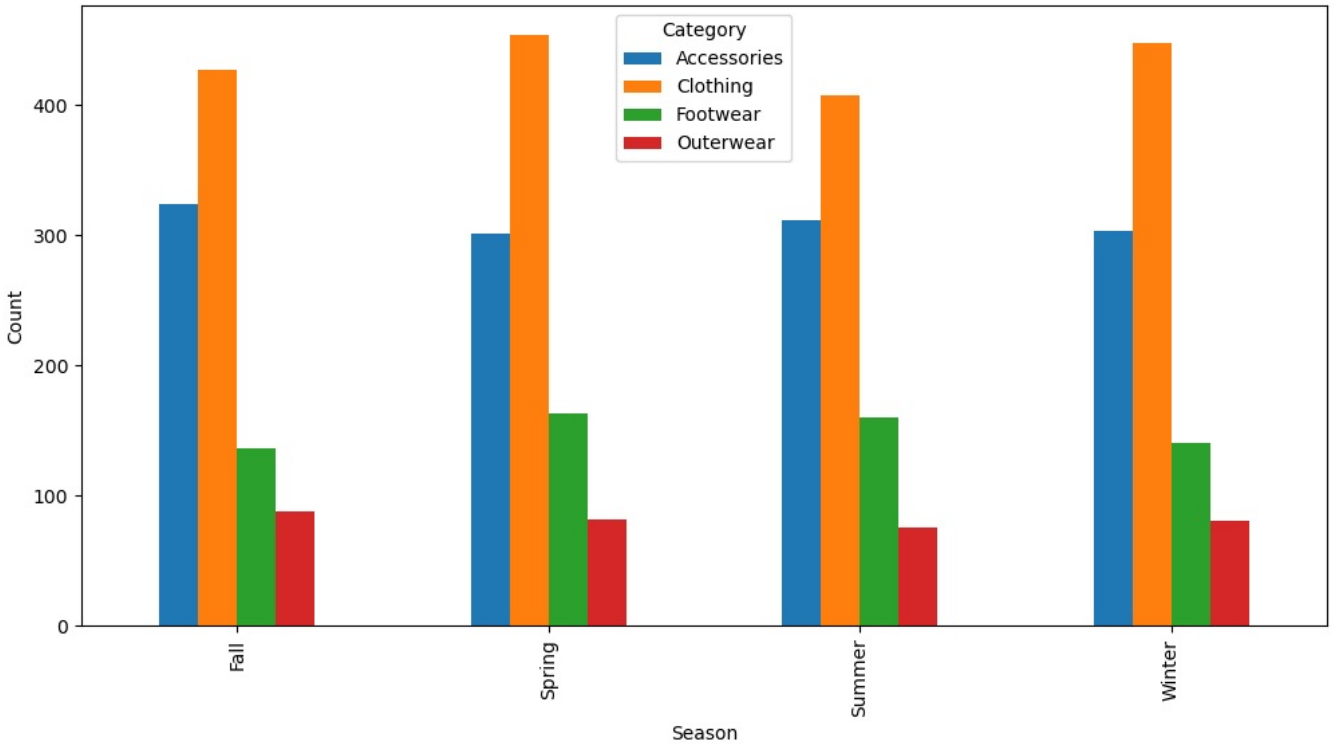
In [54]:
```
# Number of Categories that are purchased in each season

pd.crosstab(df['Season'],df['Category'])
```

Out[54]:

Category	Accessories	Clothing	Footwear	Outerwear
Season				
Fall	324	427	136	88
Spring	301	454	163	81
Summer	312	408	160	75
Winter	303	448	140	80

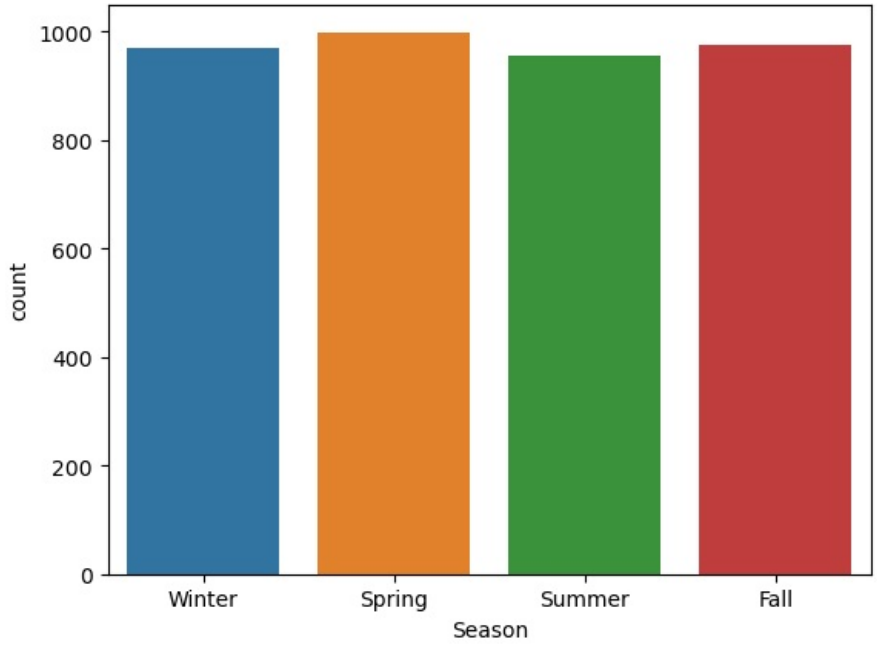
In [55]:
```
pd.crosstab(df['Season'],df['Category']).plot(kind='bar',figsize=(12,6),ylabel='Count')
```

Out[55]:
```
<Axes: xlabel='Season', ylabel='Count'>
```



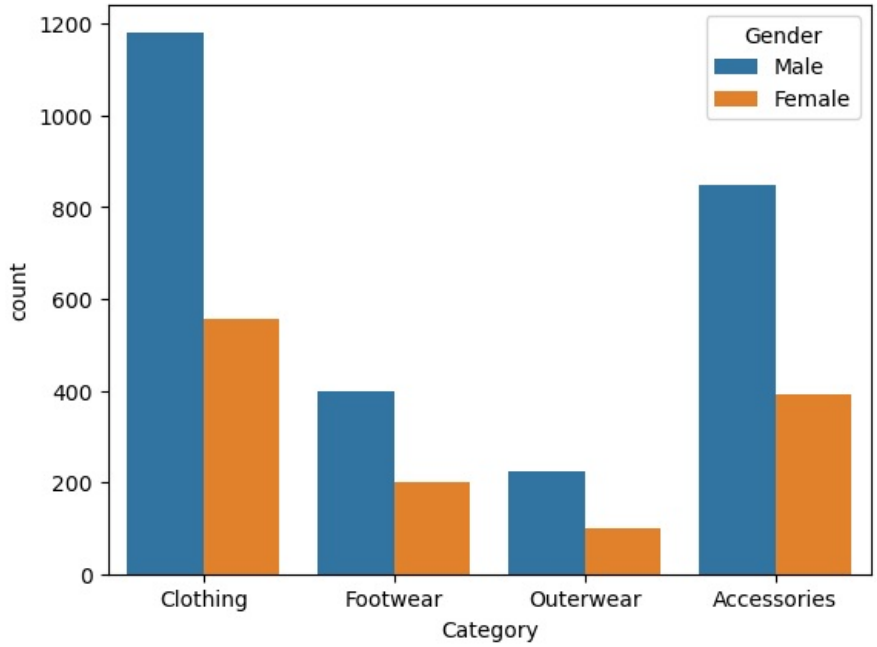
In [56]:
```
sns.countplot(data= df, x='Season')
```

Out[56]:
```
<Axes: xlabel='Season', ylabel='count'>
```



In [57]:
```
sns.countplot(data= df, x='Category',hue='Gender')
```

Out[57]:
```
<Axes: xlabel='Category', ylabel='count'>
```



In [69]:
```
import pandas as pd
def summary(data):
    print(f'data shape : {df.shape}')
    sum=pd.DataFrame(df.dtypes,columns=['data type'])
    sum["Missing"]=df.isnull().sum()
    sum["%Missing"]=(df.isnull().sum()/len(df))*100
    sum['#unique']=df.nunique().values
    desc=pd.DataFrame(df.describe(include="all").transpose())
    sum['min']=desc['min'].values
    sum['max']=desc['max'].values
    sum['first value']=df.loc[0].values
    sum['second value']=df.loc[1].values
    sum['Third value']=df.loc[2].values


    return sum
```

In [70]:
```
summary(df)
```

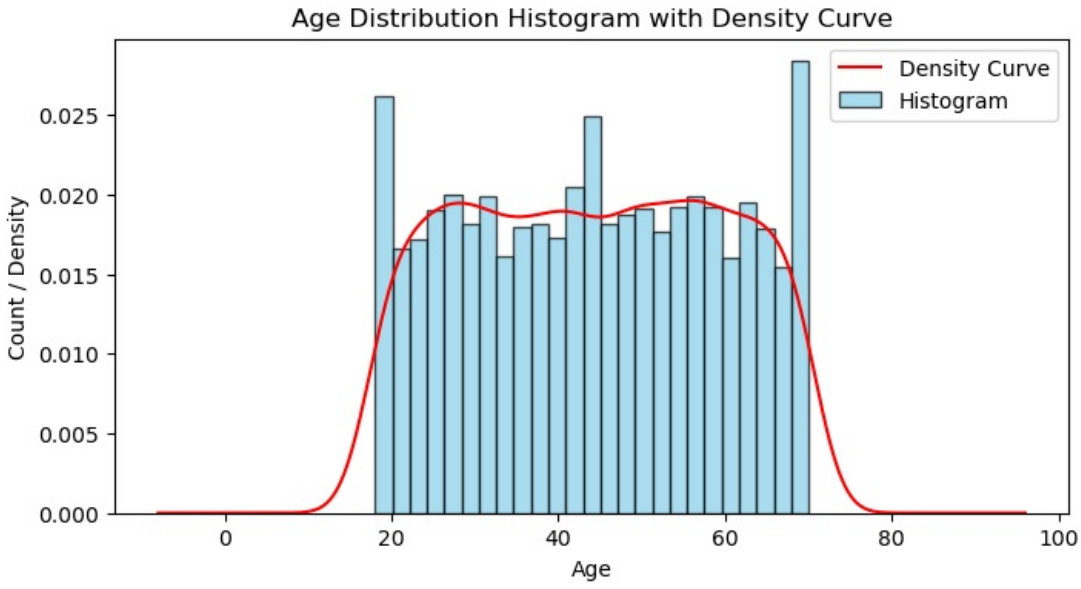
```
data shape : (3900, 19)
```

Out[70]:

	data type	Missing	%Missing	#unique	min	max	first value	second value	Third value
Customer ID	int64	0	0.0	3900	1.0	3900.0	1	2	3
Age	int64	0	0.0	53	18.0	70.0	55	19	50
Gender	object	0	0.0	2	NaN	NaN	Male	Male	Male
Item Purchased	object	0	0.0	25	NaN	NaN	Blouse	Sweater	Jeans
Category	object	0	0.0	4	NaN	NaN	Clothing	Clothing	Clothing
Purchase Amount (USD)	int64	0	0.0	81	20.0	100.0	53	64	73
Location	object	0	0.0	50	NaN	NaN	Kentucky	Maine	Massachusetts
Size	object	0	0.0	4	NaN	NaN	L	L	S
Color	object	0	0.0	25	NaN	NaN	Gray	Maroon	Maroon
Season	object	0	0.0	4	NaN	NaN	Winter	Winter	Spring
Review Rating	float64	0	0.0	26	2.5	5.0	3.1	3.1	3.1
Subscription Status	object	0	0.0	2	NaN	NaN	Yes	Yes	Yes
Payment Method	object	0	0.0	6	NaN	NaN	Credit Card	Bank Transfer	Cash
Shipping Type	object	0	0.0	6	NaN	NaN	Express	Express	Free Shipping
Discount Applied	object	0	0.0	2	NaN	NaN	Yes	Yes	Yes
Promo Code Used	object	0	0.0	2	NaN	NaN	Yes	Yes	Yes
Previous Purchases	int64	0	0.0	50	1.0	50.0	14	2	23
Preferred Payment Method	object	0	0.0	6	NaN	NaN	Venmo	Cash	Credit Card
Frequency of Purchases	object	0	0.0	7	NaN	NaN	Fortnightly	Fortnightly	Weekly

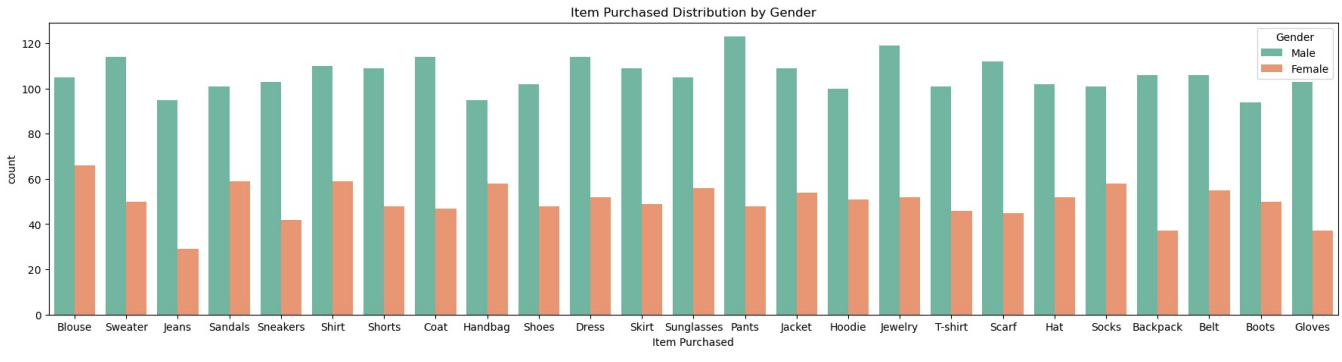
In [71]:

```
plt.figure(figsize=(8, 4))
plt.hist(df['Age'],edgecolor = 'black',alpha=0.7,bins=25,color = 'skyblue',density=True)
df['Age'].plot(kind='kde', color = 'red')
plt.xlabel('Age')
plt.ylabel('Count / Density')
plt.title('Age Distribution Histogram with Density Curve')
plt.legend(['Density Curve', 'Histogram'])
plt.show()
```

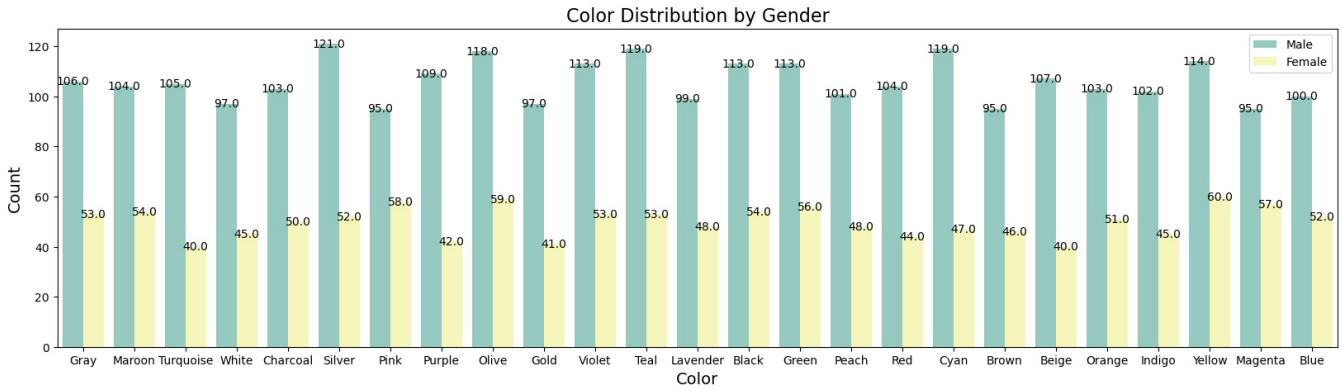


In [73]:

```
plt.figure(figsize=(22,5))
sns.countplot(data=df,x='Item Purchased',hue='Gender',palette='Set2')
plt.title('Item Purchased Distribution by Gender')
plt.show()
```

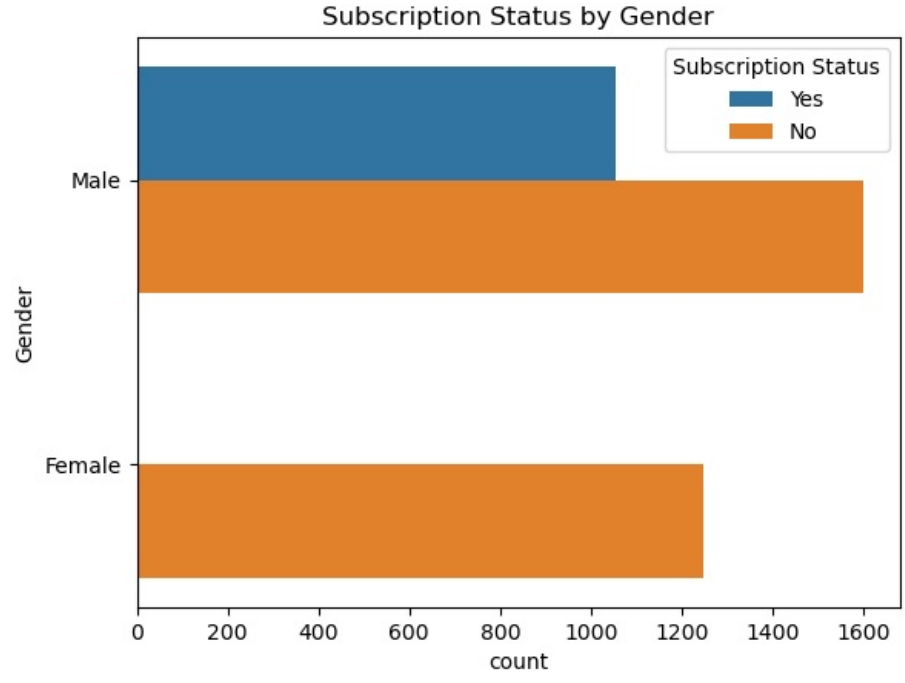


In [79]:
```
plt.figure(figsize=(20,5))
ax=sns.countplot(data = df,x='Color',hue='Gender',palette='Set3')
for P in ax.patches:
    ax.annotate(f'{P.get_height()}', (P.get_x() + P.get_width() / 2., P.get_height()), ha='center', va='center'
    plt.xlabel('Color', fontsize=14)
    plt.ylabel('Count', fontsize=14)
    plt.title('Color Distribution by Gender', fontsize=16)
    plt.legend(title='Gender', fontsize=8, title_fontsize=8)
    plt.legend()
```



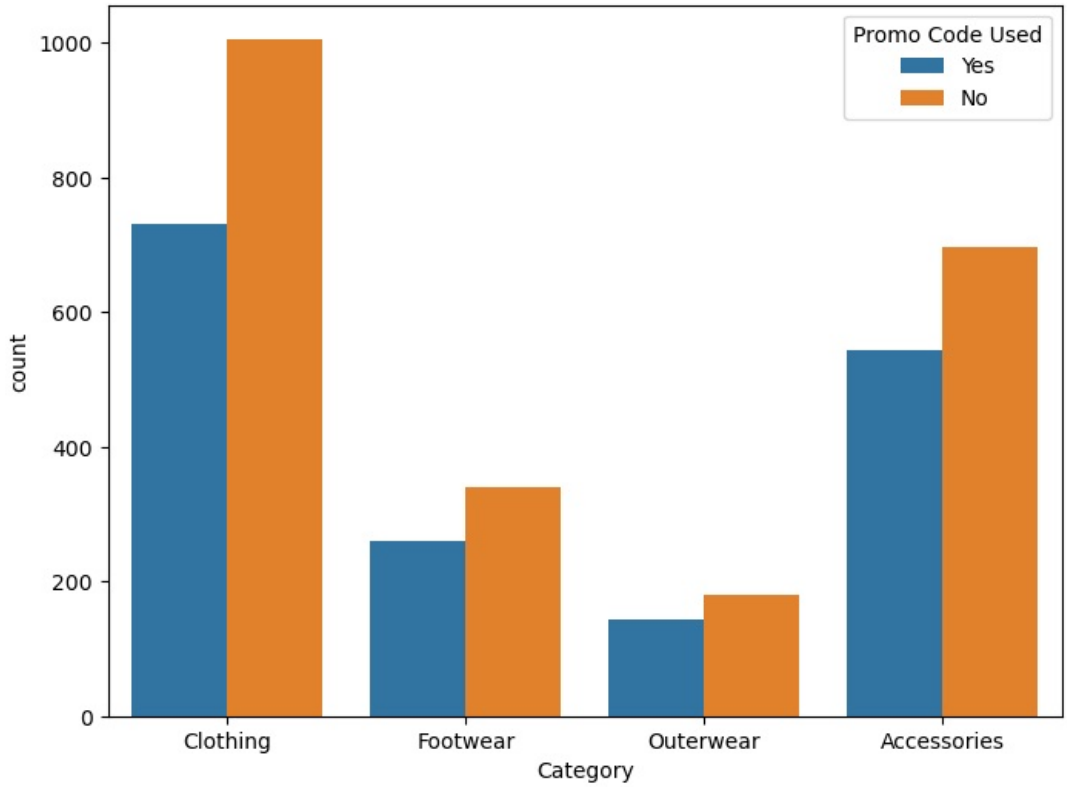
In [80]:
```
plt.figure(figsize=(20, 6))
sns.violinplot(x='Category', y='Review Rating', data=df)
plt.title('Review Rating Category')
plt.xlabel('category')
plt.ylabel('Review Rating')
plt.xticks(rotation=45,fontsize=12)
plt.show()
```

In [83]:
```
sns.countplot(data=df,y='Gender',hue='Subscription Status')
plt.title('Subscription Status by Gender')
plt.show()
```



In [84]:
```
plt.figure(figsize=(8,6))
sns.countplot(data=df,x='Category',hue='Promo Code Used')
```

Out[84]:
```
<Axes: xlabel='Category', ylabel='count'>
```



In [ ]:

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js