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

In [2]:
```
data= pd.read_csv(r"H:\Data\Prodigy Infotech\PURCHASE.CSV")
```

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

Out[3]:

	User_ID	Product_ID	Gender	Age	City_Category	Stay_In_Current_City_Years	Marital_Stat
0	1000001	P00069042	F	0-17	A	2	
1	1000001	P00248942	F	0-17	A	2	
2	1000001	P00087842	F	0-17	A	2	
3	1000001	P00085442	F	0-17	A	2	
4	1000002	P00285442	M	55+	C	4+	

In [4]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 263015 entries, 0 to 263014
Data columns (total 11 columns):
 #   Column                      Non-Null Count   Dtype  
---  ------                      --------------   -----  
 0   User_ID                     263015 non-null  int64  
 1   Product_ID                  263014 non-null  object 
 2   Gender                      263014 non-null  object 
 3   Age                         263014 non-null  object 
 4   City_Category               263014 non-null  object 
 5   Stay_In_Current_City_Years  263014 non-null  object 
 6   Marital_Status              263014 non-null  float64
 7   Product_Category_1          263014 non-null  float64
 8   Product_Category_2          181501 non-null  float64
 9   Product_Category_3          80582 non-null   float64
 10  Purchase                    263014 non-null  float64
dtypes: float64(5), int64(1), object(5)
memory usage: 22.1+ MB
```

In [5]:
```
data.columns
```

Out[5]:
```
Index(['User_ID', 'Product_ID', 'Gender', 'Age', 'City_Category',
       'Stay_In_Current_City_Years', 'Marital_Status', 'Product_Category_
1',
       'Product_Category_2', 'Product_Category_3', 'Purchase'],
      dtype='object')
```

In [6]:
```
data['Age'].nunique()
```

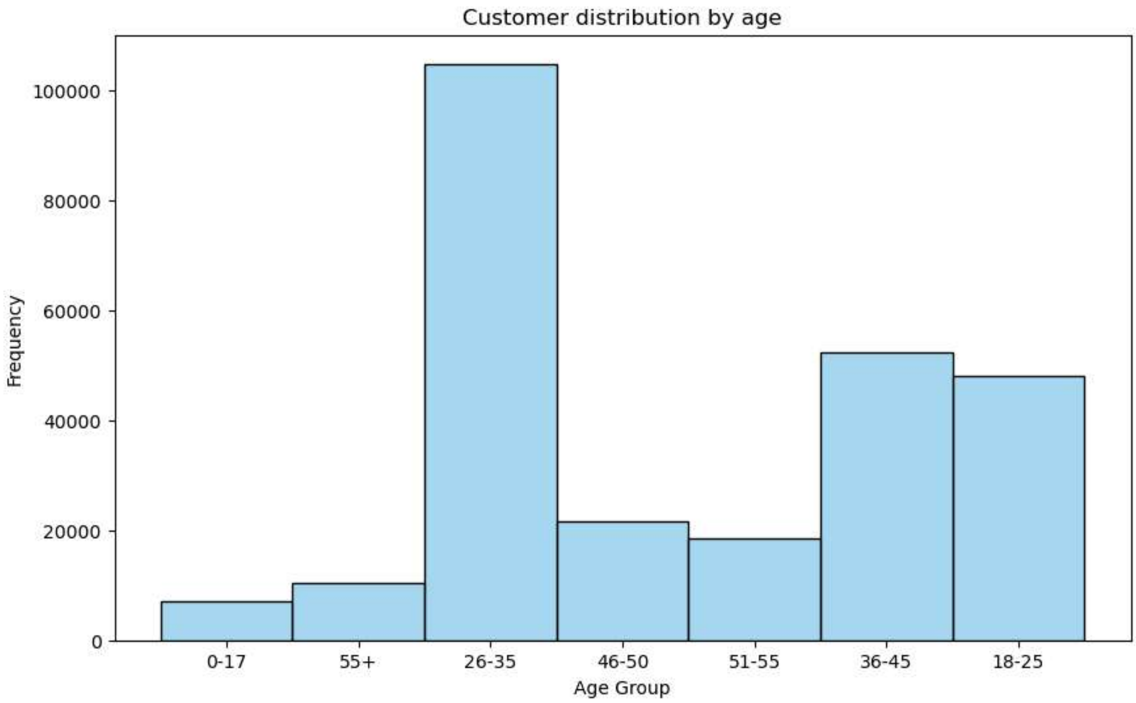
Out[6]:
```
7
```

From the below code snippet, we observe that the majority of customers fall within the 26 to 35 age range, while the 0 to 17 age range comprises the smallest customer segment. These observations are further illustrated in the bar chart below, providing a precise visual representation.

In [7]:
```
data['Age'].value_counts()
```

Out[7]:
```
26-35    104912
36-45     52396
18-25     48193
46-50     21619
51-55     18509
55+       10321
0-17       7064
Name: Age, dtype: int64
```

In [8]:
```
plt.figure(figsize= (10,6))
sns.histplot(data['Age'], bins= 7, color= 'skyblue')
plt.title('Customer distribution by age')
plt.xlabel('Age Group')
plt.ylabel('Frequency')
plt.show()
```

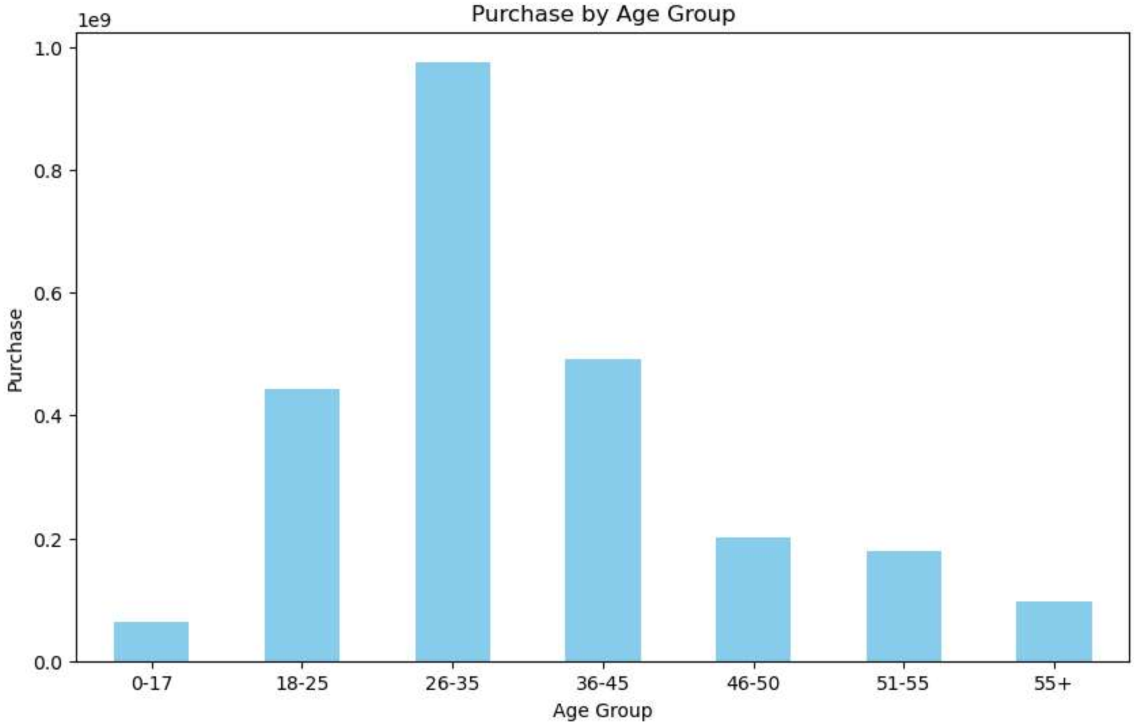


From the below code snippet, we can observe that the highest purchase amount is made by customers in the 26 to 35 age group, followed by those in the 36 to 45 age group. The lowest purchase amount is attributed to customers in the 0 to 17 age group. These observations are further illustrated in the bar chart below, providing a clearer visual representation.

In [9]:
```
purchase_data= data.groupby('Age')['Purchase'].sum()
purchase_data
```

Out[9]:
```
Age
0-17       64173683.0
18-25     442696277.0
26-35     975615086.0
36-45     492346613.0
46-50     200909949.0
51-55     178134937.0
55+        97231211.0
Name: Purchase, dtype: float64
```

In [10]:
```
purchase_data= data.groupby('Age')['Purchase'].sum()
purchase_data.plot(kind= 'bar', color= 'skyblue', figsize= (10,6))
plt.xlabel('Age Group')
plt.ylabel('Purchase')
plt.xticks(rotation= 0)
plt.title('Purchase by Age Group')
plt.show()
```

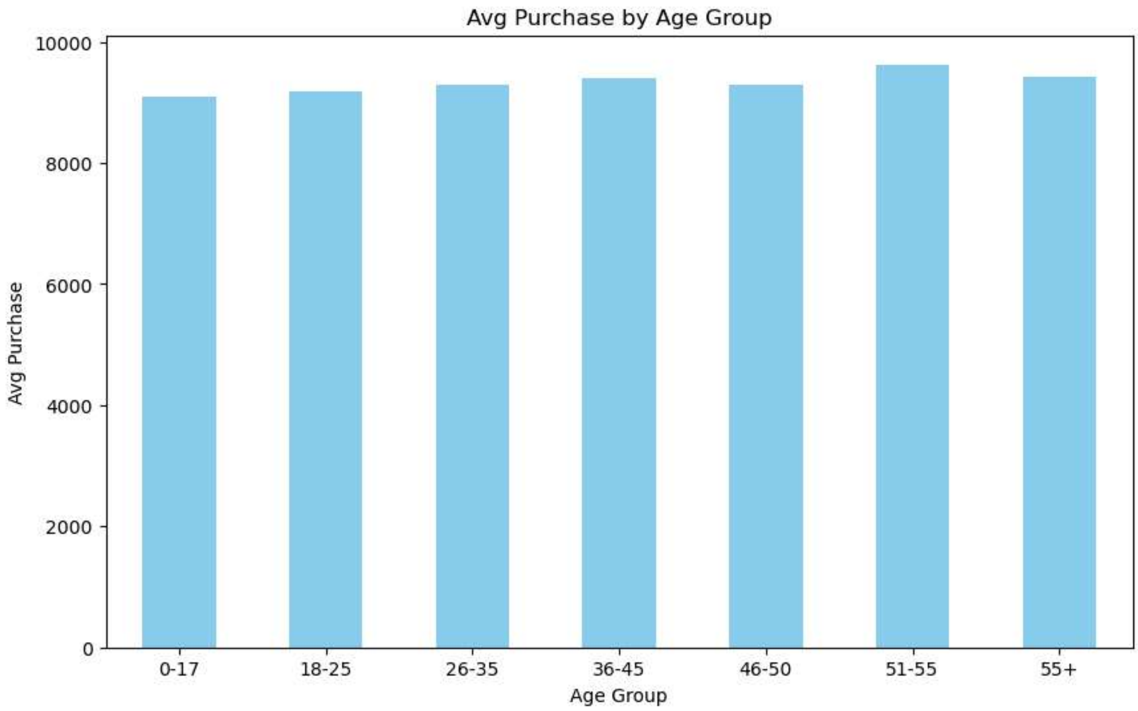


Here's the interesting part: As observed from the above code snippets, the highest total purchase amount was made by the 26 to 35 age group. However, the below code snippet reveals that the highest average purchase amount is by customers in the 51 to 55 age group. Conversely, the lowest average purchase amount remains with the 0 to 17 age group. This is more clearly illustrated in the bar chart below. Although the difference not very signigicant.

In [11]:
```
Avg_purchase_data= data.groupby('Age')['Purchase'].mean()
Avg_purchase_data
```

Out[11]:
```
Age
0-17     9084.609711
18-25    9185.904115
26-35    9299.366002
36-45    9396.645030
46-50    9293.211943
51-55    9624.233454
55+      9420.716113
Name: Purchase, dtype: float64
```

In [12]:
```
Avg_purchase_data.plot(kind= 'bar', color= 'skyblue', figsize=(10,6))
plt.xlabel('Age Group')
plt.ylabel('Avg Purchase')
plt.title('Avg Purchase by Age Group')
plt.xticks(rotation=0)
plt.show()
```



From the code snippet below, we can clearly observe that male customers have made a significantly higher amount of purchases. The bar graph provided offers a better visual representation of this observation.

In [13]:
```
gender_purchase= data.groupby('Gender')['Purchase'].sum()
gender_purchase
```

Out[13]:
```
Gender
F    5.681118e+08
M    1.882996e+09
Name: Purchase, dtype: float64
```

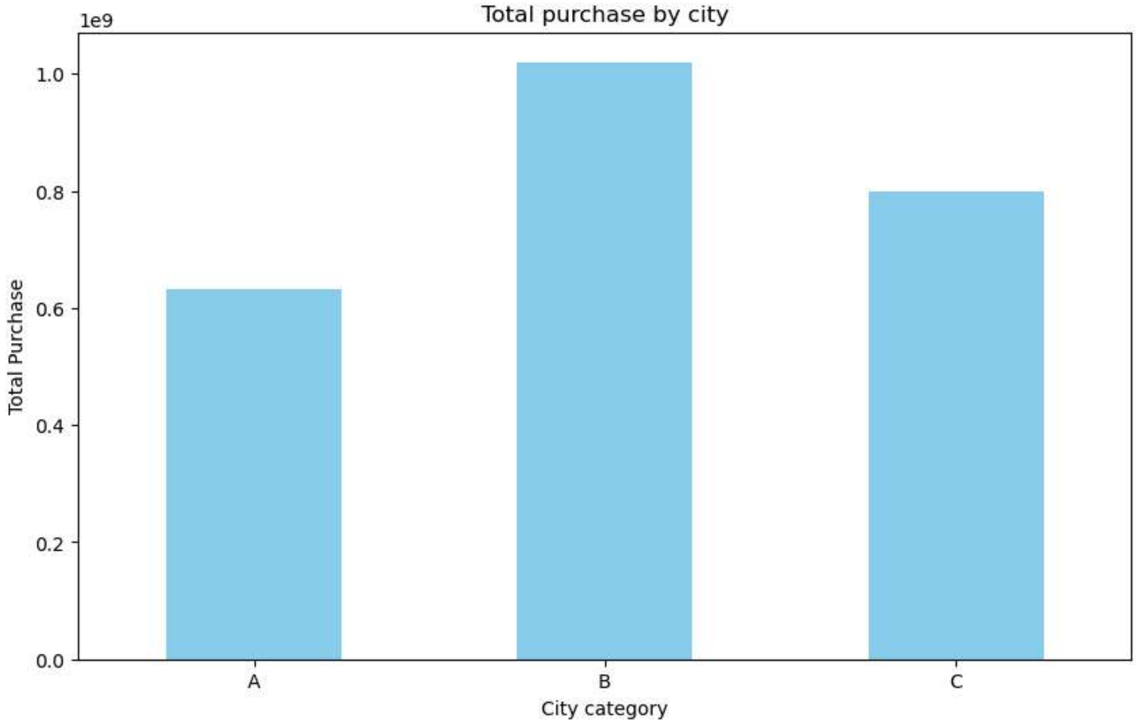
In [14]:
```
gender_purchase.plot(kind= 'bar', color= 'skyblue', figsize=(10,6))
plt.xlabel('Gender')
plt.ylabel('Purchase')
plt.title('Total Purchase by Gender')
plt.xticks(rotation=0)
plt.show()
```

From the code snippet below, we can observe that customers residing in city category B have spent significantly higher amounts compared to those in the other two city categories. This observation is better visualized in the bar graph below.

In [15]:
```
purchase_by_city= data.groupby('City_Category')['Purchase'].sum()
purchase_by_city
```

Out[15]:
```
City_Category
A    6.327101e+08
B    1.019965e+09
C    7.984328e+08
Name: Purchase, dtype: float64
```

In [16]:
```
purchase_by_city.plot(kind='bar', color='skyblue', figsize=(10,6))
plt.xlabel('City category')
plt.ylabel('Total Purchase')
plt.title('Total purchase by city')
plt.xticks(rotation=0)
plt.show()
```

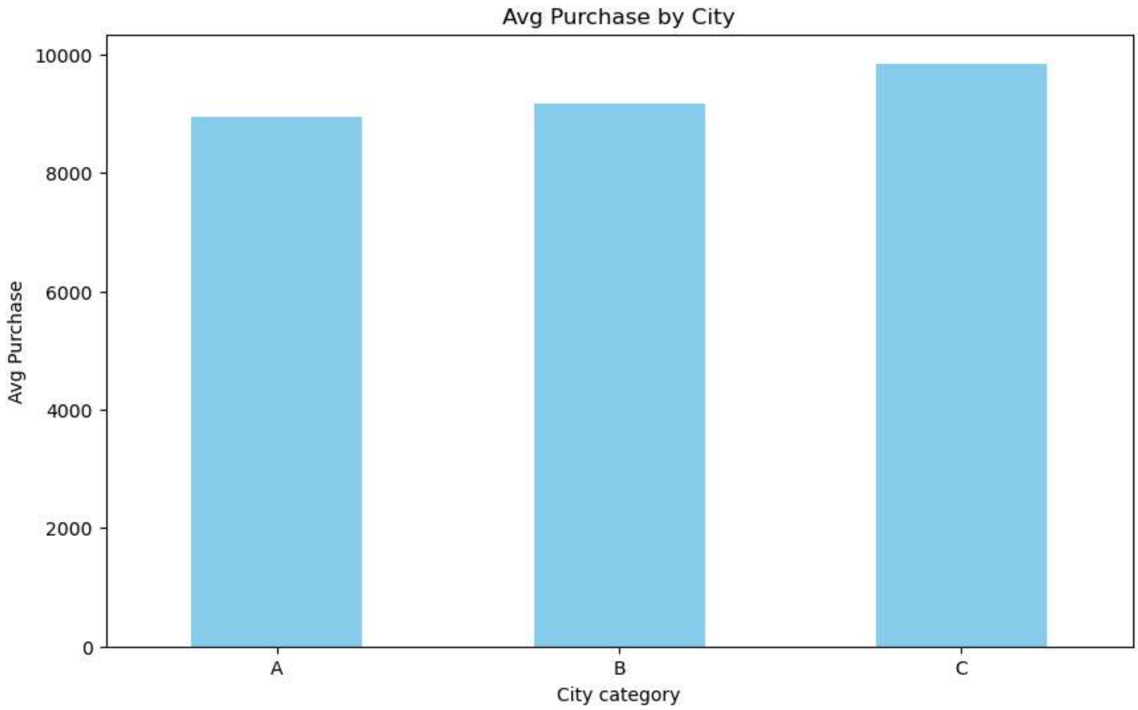


Interestingly, customers living in city category C rank highest when we compare the average purchase amounts made by customers. The bar graph provided below offers a clearer visualization of this trend.

In [17]:
```
Avg_purchase_city= data.groupby('City_Category')['Purchase'].mean()
Avg_purchase_city
```

Out[17]:
```
City_Category
A    8937.845402
B    9178.618993
C    9845.040974
Name: Purchase, dtype: float64
```

In [18]:
```
Avg_purchase_city.plot(kind='bar', color='skyblue', figsize=(10,6))
plt.xlabel('City category')
plt.ylabel('Avg Purchase')
plt.title('Avg Purchase by City')
plt.xticks(rotation=0)
plt.show()
```

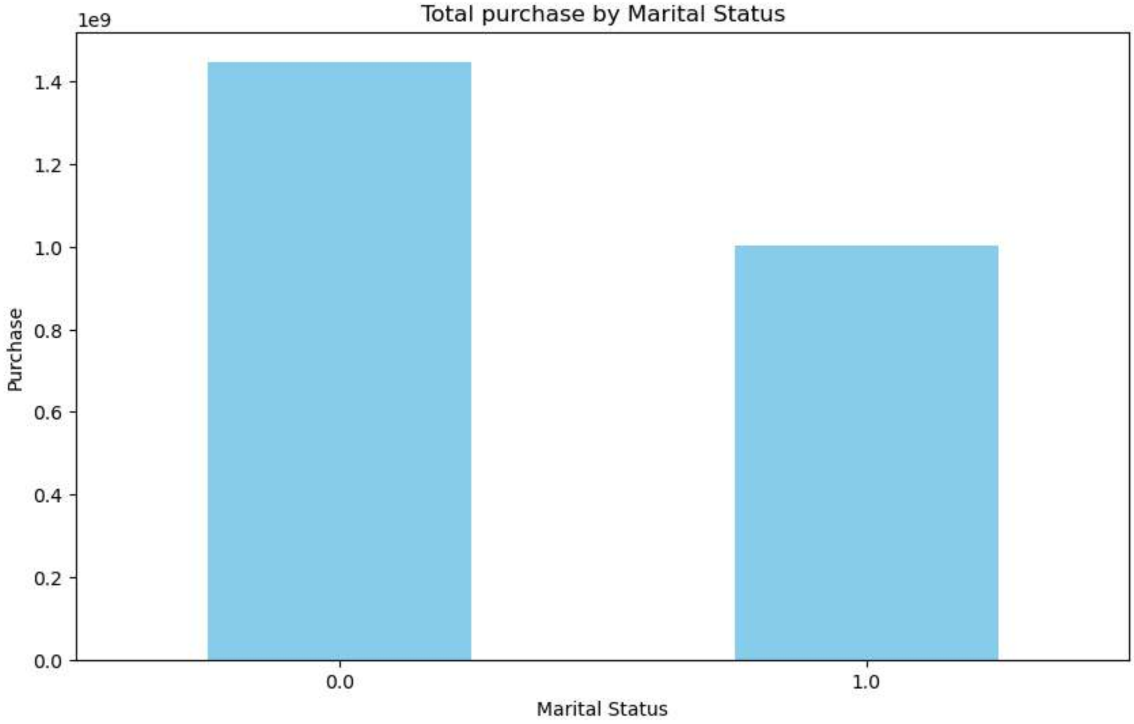


From the code snippet below, we can observe that unmarried customers have spent more compared to married customers.

In [19]:
```
purchase_by_maritalstatus= data.groupby('Marital_Status')['Purchase'].sum()
purchase_by_maritalstatus
```

Out[19]:
```
Marital_Status
0.0    1.447350e+09
1.0    1.003758e+09
Name: Purchase, dtype: float64
```

In [20]:
```
purchase_by_maritalstatus.plot(kind= 'bar', color= 'skyblue', figsize=(10,6
plt.title('Total purchase by Marital Status')
plt.xlabel('Marital Status')
plt.ylabel('Purchase')
plt.xticks(rotation=0)
plt.show()
```

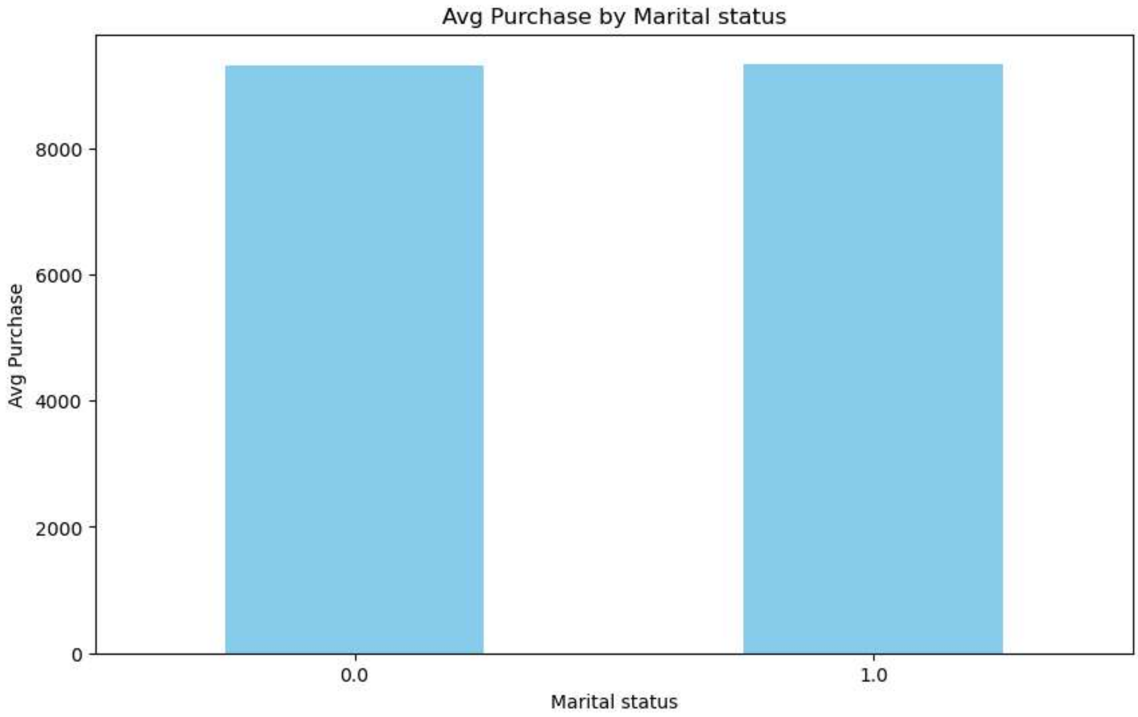


On comparing the average amount spent by customers, it's evident that married customers have, on average, spent more than unmarried customers. The bar graph provided below offers a clearer visualization of this comparison.

In [21]:
```
Avg_purchase_by_maritalstatus= data.groupby('Marital_Status')['Purchase'].m
Avg_purchase_by_maritalstatus
```

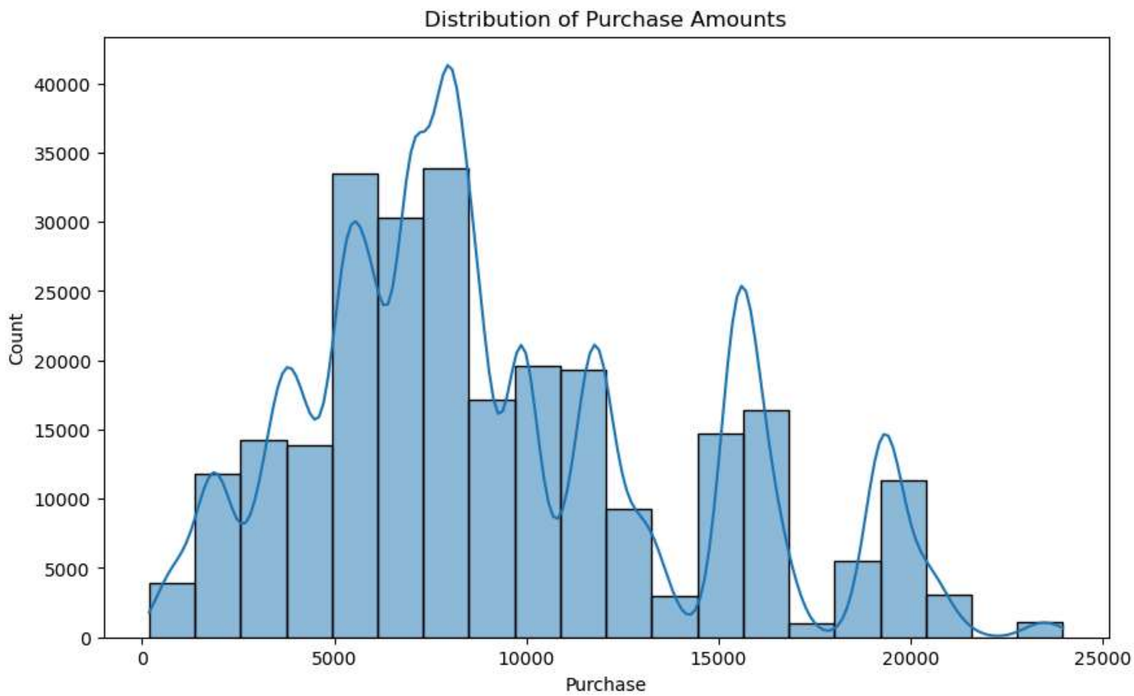
Out[21]:
```
Marital_Status
0.0    9306.279719
1.0    9338.151540
Name: Purchase, dtype: float64
```

In [22]:
```
Avg_purchase_by_maritalstatus.plot(kind='bar', color= 'skyblue', figsize=(1(
plt.title('Avg Purchase by Marital status')
plt.xlabel('Marital status')
plt.ylabel('Avg Purchase')
plt.xticks(rotation=0)
plt.show()
```



From the histogram provided below, it's evident that the majority of shopping transactions fall within the range of 5000 to 8000 purchase amount.

In [23]:
```
plt.figure(figsize=(10, 6))
sns.histplot(data['Purchase'], kde=True, bins=20)
plt.title('Distribution of Purchase Amounts')
plt.show()
```

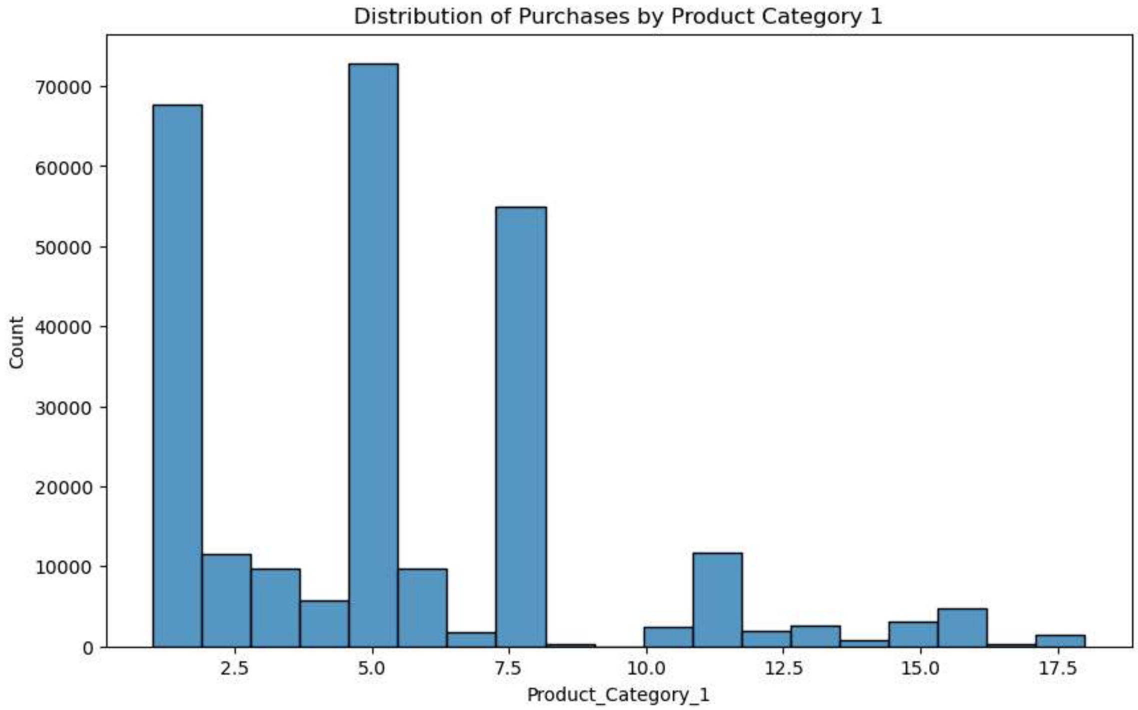


From the histogram provided below, it's apparent that product 5.0 in product category 1 is the highest selling product, followed by product 1.0, with product 8.0 ranking third in terms of sales volume.

In [24]:
```
data['Product_Category_1'].value_counts()
```

Out[24]:
```
5.0     72899
1.0     67659
8.0     54987
11.0    11685
2.0     11528
3.0      9729
6.0      9684
4.0      5681
16.0     4727
15.0     3025
13.0     2636
10.0     2436
12.0     1871
7.0      1782
18.0     1481
14.0      722
17.0      289
9.0       193
Name: Product_Category_1, dtype: int64
```

In [25]:
```
plt.figure(figsize=(10, 6))
sns.histplot(data['Product_Category_1'], kde=False, bins=len(data['Product_(
plt.title('Distribution of Purchases by Product Category 1')
plt.show()
```

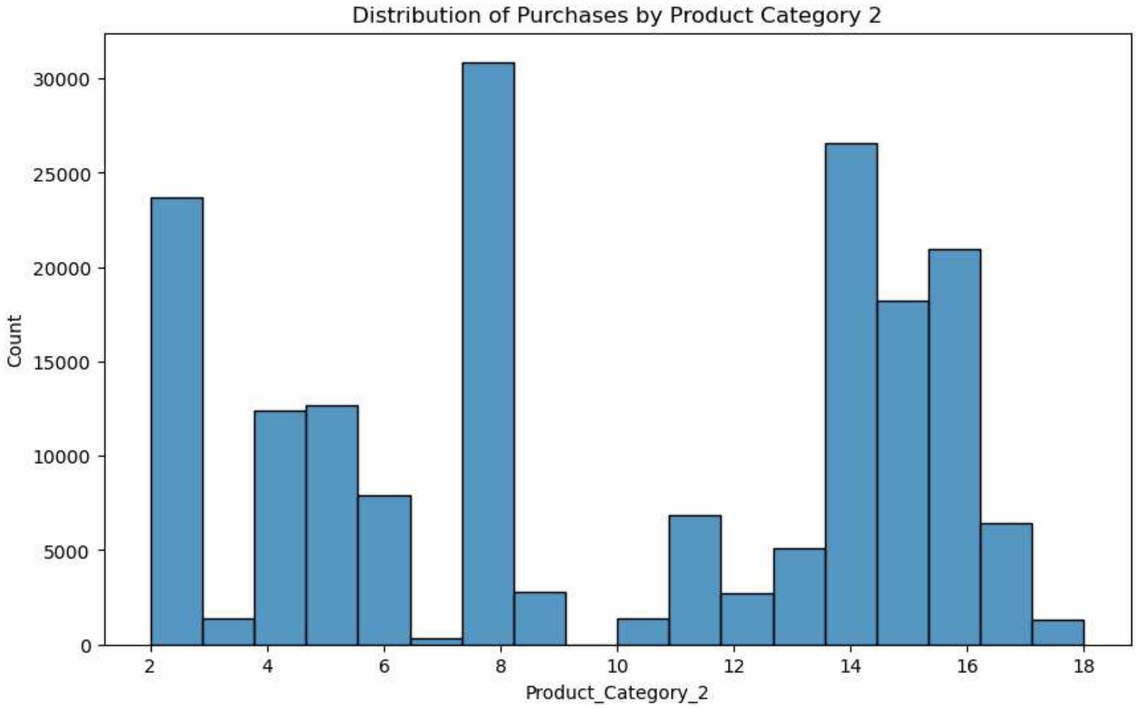


Based on the Series provided below and the accompanying histogram, it is evident that within product category 2, the top-selling products are 8.0, followed by 14.0, with product 2.0 ranking third in terms of sales volume.

In [26]:
```
data['Product_Category_2'].value_counts()
```

Out[26]:
```
8.0     30889
14.0    26570
2.0     23667
16.0    20935
15.0    18240
5.0     12659
4.0     12427
6.0      7937
11.0     6836
17.0     6412
13.0     5079
9.0      2754
12.0     2675
3.0      1403
10.0     1385
18.0     1329
7.0       304
Name: Product_Category_2, dtype: int64
```

In [27]:
```
plt.figure(figsize=(10, 6))
sns.histplot(data['Product_Category_2'], kde=False, bins=len(data['Product_(
plt.title('Distribution of Purchases by Product Category 2')
plt.show()
```

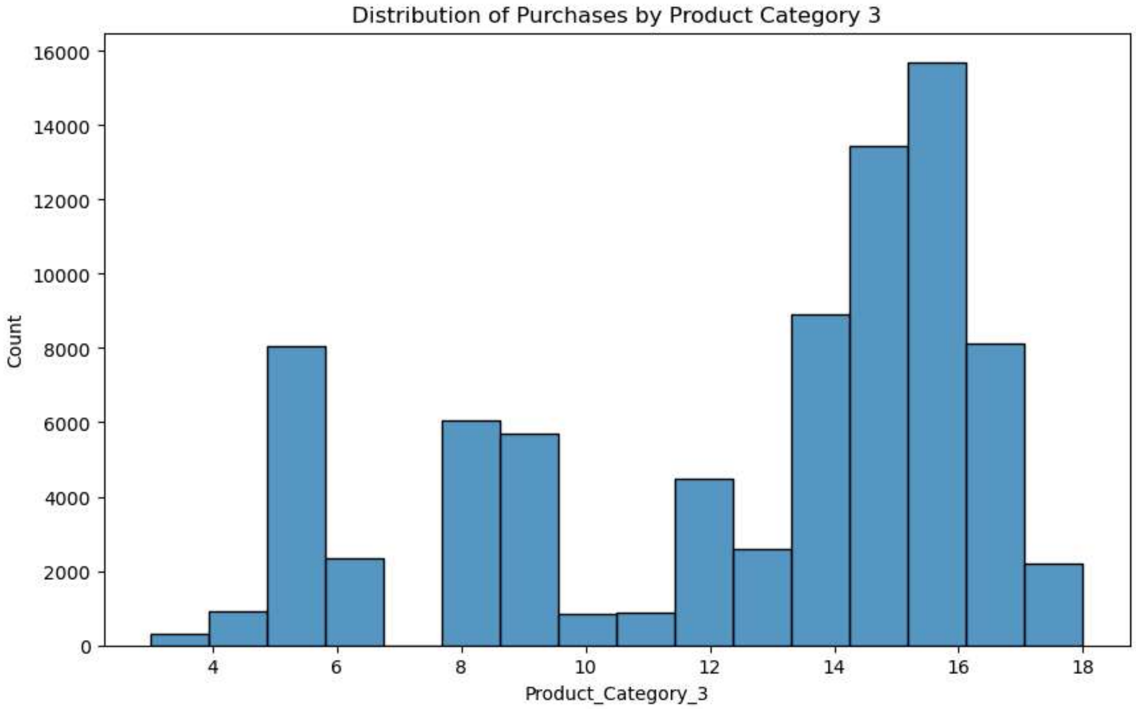


Based on the Series provided below and the accompanying histogram, it is evident that within product category 3, the top-selling products are 16.0, followed by 15.0, with product 14.0 ranking third in terms of sales volume.

In [28]:
```
data['Product_Category_3'].value_counts()
```

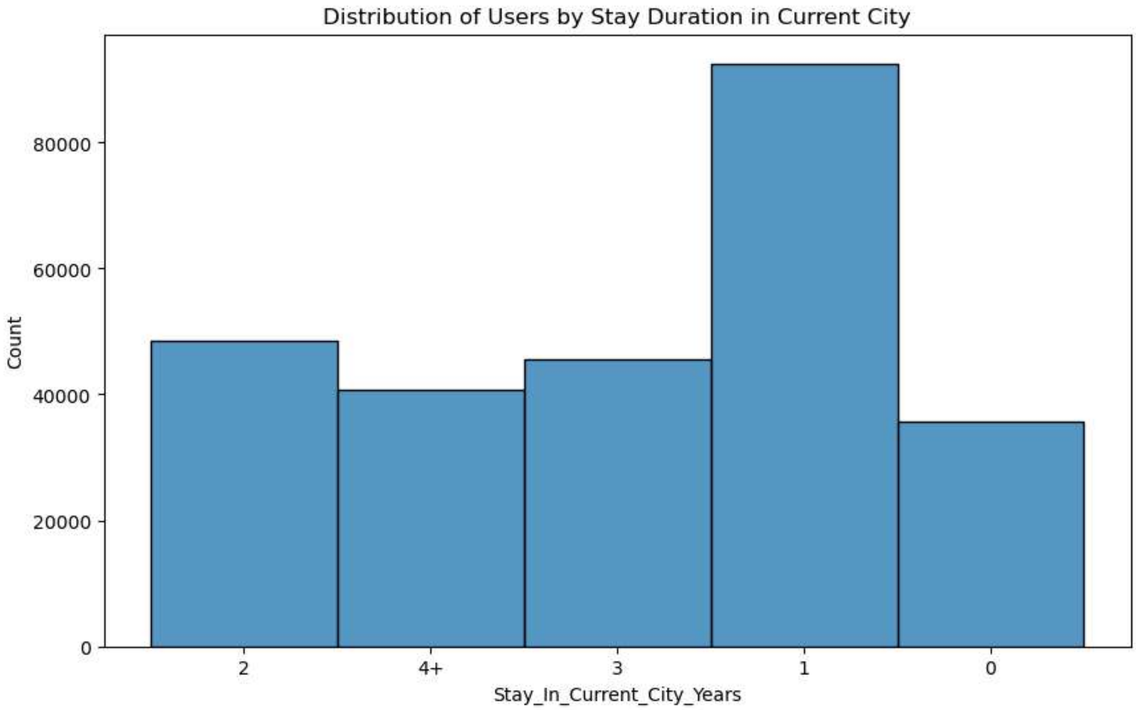
Out[28]:
```
16.0    15706
15.0    13448
14.0     8908
17.0     8135
5.0      8068
8.0      6064
9.0      5691
12.0     4479
13.0     2584
6.0      2349
18.0     2213
4.0       918
11.0      893
10.0      828
3.0       298
Name: Product_Category_3, dtype: int64
```

In [29]:
```
plt.figure(figsize=(10, 6))
sns.histplot(data['Product_Category_3'], kde=False, bins=len(data['Product_(
plt.title('Distribution of Purchases by Product Category 3')
plt.show()
```



From the below histogram it is evident that the maximum number of customers are the one's who are living in the current city for 1 year.

In [30]:
```
plt.figure(figsize=(10, 6))
sns.histplot(data['Stay_In_Current_City_Years'], kde=False, bins=len(data['
plt.title('Distribution of Users by Stay Duration in Current City')
plt.show()
```



Upon examining the provided DataFrame and the accompanying combined bar graph, it is apparent that female customers have, on average, spent less across all age groups compared to male customers. Specifically, the lowest average spending is observed among females aged 18 to 25, whereas the highest average spending has been done by males of age between 51 to 55.

In [31]:
```
purchase_age_gender= data.groupby(['Age', 'Gender'])['Purchase'].mean().rese
purchase_age_gender
```

Out[31]:

	Age	Gender	Purchase
0	0-17	F	8559.298745
1	0-17	M	9353.221866
2	18-25	F	8406.001602
3	18-25	M	9440.512276
4	26-35	F	8795.096858
5	26-35	M	9448.392541
6	36-45	F	9000.520879
7	36-45	M	9525.175101
8	46-50	F	8917.703633
9	46-50	M	9446.812292
10	51-55	F	9177.374608
11	51-55	M	9780.210641
12	55+	F	9016.633497
13	55+	M	9545.956593

In [32]:
```
sns.barplot(data= purchase_age_gender, x='Age', y='Purchase', hue= 'Gender'
plt.title('Avg Purchase amount by Age and Gender')
plt.xlabel('Age, Gender')
plt.ylabel('Avg Purchase')
plt.show()
```