In [100]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import os
from matplotlib import pyplot
```

In [101]:
```
ap = pd.read_csv("/Users/anandchauhan/Downloads/Blackfriday.csv")
ap.head(10)
```

Out[101]:

	User_ID	Product_ID	Gender	Age	Occupation	City_Category	Stay_In_Current_City_Years	Ma
0	1000001	P00069042	F	0-17	10	A	2	
1	1000001	P00248942	F	0-17	10	A	2	
2	1000001	P00087842	F	0-17	10	A	2	
3	1000001	P00085442	F	0-17	10	A	2	
4	1000002	P00285442	M	55+	16	C	4+	
5	1000003	P00193542	M	26-35	15	A	3	
6	1000004	P00184942	M	46-50	7	B	2	
7	1000004	P00346142	M	46-50	7	B	2	
8	1000004	P0097242	M	46-50	7	B	2	
9	1000005	P00274942	M	26-35	20	A	1	

In [102]:
```
ap.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 537577 entries, 0 to 537576
Data columns (total 12 columns):
User_ID                     537577 non-null int64
Product_ID                  537577 non-null object
Gender                      537577 non-null object
Age                         537577 non-null object
Occupation                  537577 non-null int64
City_Category               537577 non-null object
Stay_In_Current_City_Years  537577 non-null object
Marital_Status              537577 non-null int64
Product_Category_1          537577 non-null int64
Product_Category_2          370591 non-null float64
Product_Category_3          164278 non-null float64
Purchase                    537577 non-null int64
dtypes: float64(2), int64(5), object(5)
memory usage: 49.2+ MB
```

In [103]:
```
ap.isnull().sum()
```

Out[103]:
```
User_ID                          0
Product_ID                       0
Gender                           0
Age                              0
Occupation                       0
City_Category                    0
Stay_In_Current_City_Years       0
Marital_Status                   0
Product_Category_1               0
Product_Category_2          166986
Product_Category_3          373299
Purchase                         0
dtype: int64
```

In [104]:
```
ap.columns
```

Out[104]:
```
Index(['User_ID', 'Product_ID', 'Gender', 'Age', 'Occupation', 'City_C
ategory',
       'Stay_In_Current_City_Years', 'Marital_Status', 'Product_Catego
ry_1',
       'Product_Category_2', 'Product_Category_3', 'Purchase'],
      dtype='object')
```

In [110]:
```
ap.sort_values('User_ID').head(10)
#ap['User_ID'].value_counts().count()
```

Out[110]:

	User_ID	Product_ID	Gender	Age	Occupation	City_Category	Stay_In_Current_City_Year
0	1000001	P00069042	F	0-17	10	A	
390151	1000001	P00255842	F	0-17	10	A	
390150	1000001	P0097142	F	0-17	10	A	
350797	1000001	P00289942	F	0-17	10	A	
311713	1000001	P00210342	F	0-17	10	A	
311712	1000001	P00248442	F	0-17	10	A	
311711	1000001	P00051442	F	0-17	10	A	
311710	1000001	P00183942	F	0-17	10	A	
311709	1000001	P00178342	F	0-17	10	A	
467663	1000001	P00058142	F	0-17	10	A	

In [111]:
```
ap['User_ID'].value_counts().count()
```

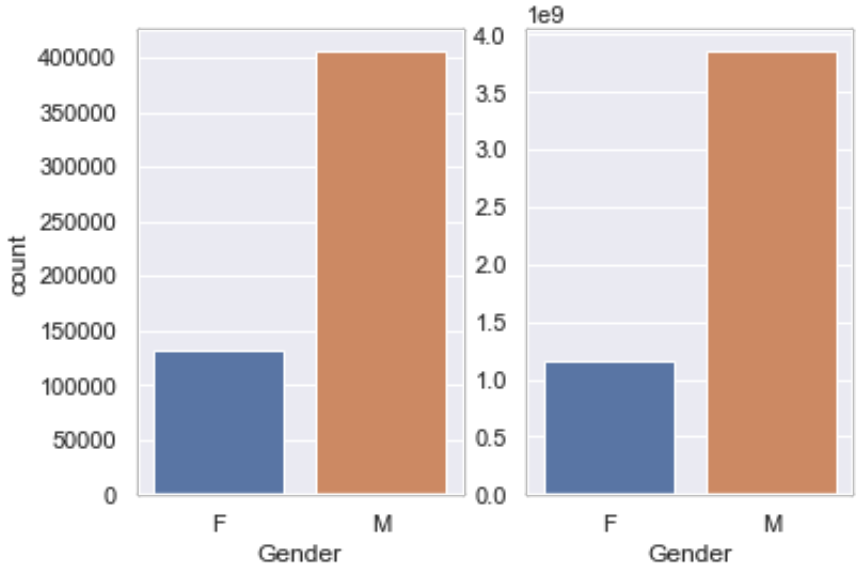
Out[111]: 5891

In [112]:
```
ap['Gender'].unique()
```

Out[112]: array(['F', 'M'], dtype=object)

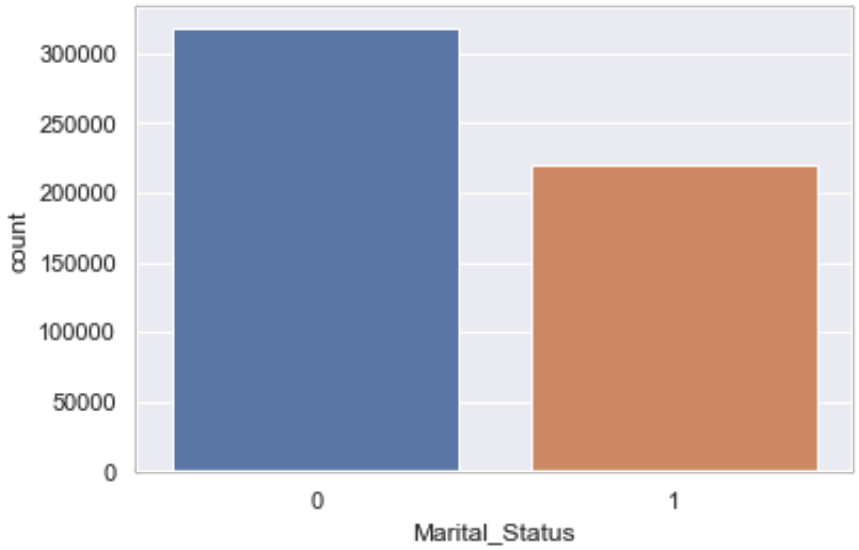
In [113]:
```
plt.subplot(1,2,1)
sns.countplot(ap['Gender']) #attendance

m_purchase = ap.groupby(['Gender'])['Purchase'].sum()
plt.subplot(1,2,2)
sns.barplot(m_purchase.index, m_purchase.values) #dollar value
```

Out[113]: <matplotlib.axes._subplots.AxesSubplot at 0x1a2f7b9630>



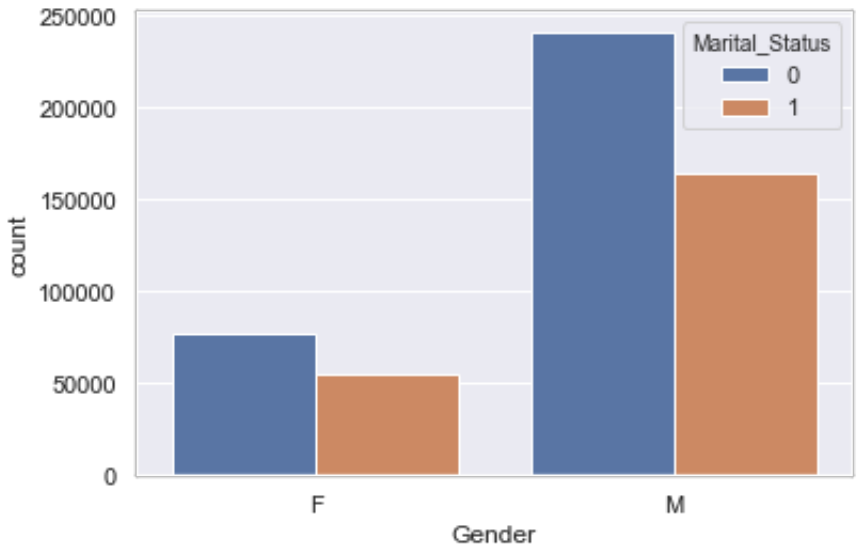
In [114]:
```
sns.countplot(ap['Marital_Status'])
```

Out[114]: <matplotlib.axes._subplots.AxesSubplot at 0x1a2aa27b00>



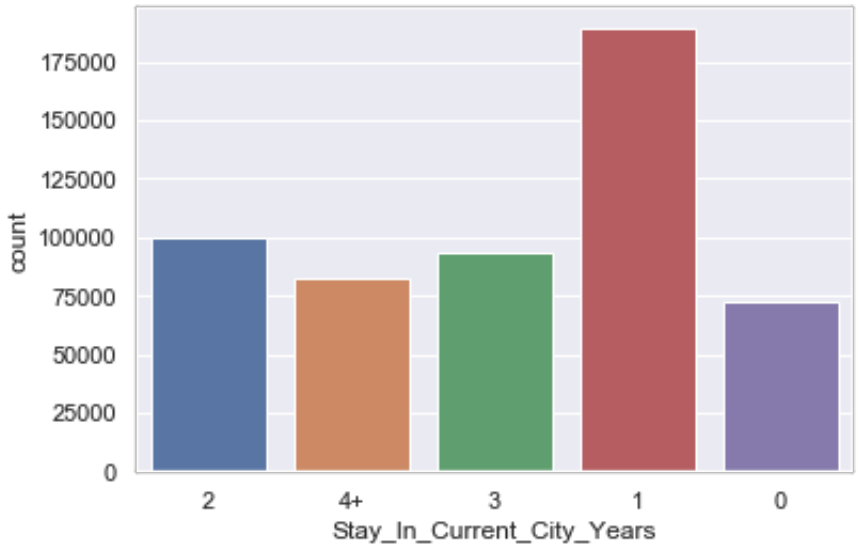
In [115]:
```
sns.countplot(ap['Gender'], hue = ap['Marital_Status'])
```

Out[115]: <matplotlib.axes._subplots.AxesSubplot at 0x1a16bc5e80>



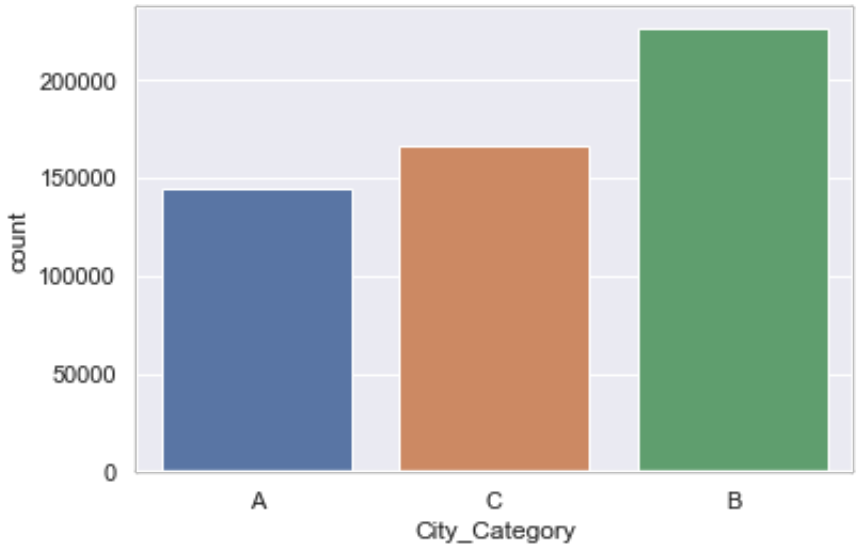
In [116]:
```
sns.countplot(ap['Stay_In_Current_City_Years'])
```

Out[116]: <matplotlib.axes._subplots.AxesSubplot at 0x1a32272630>



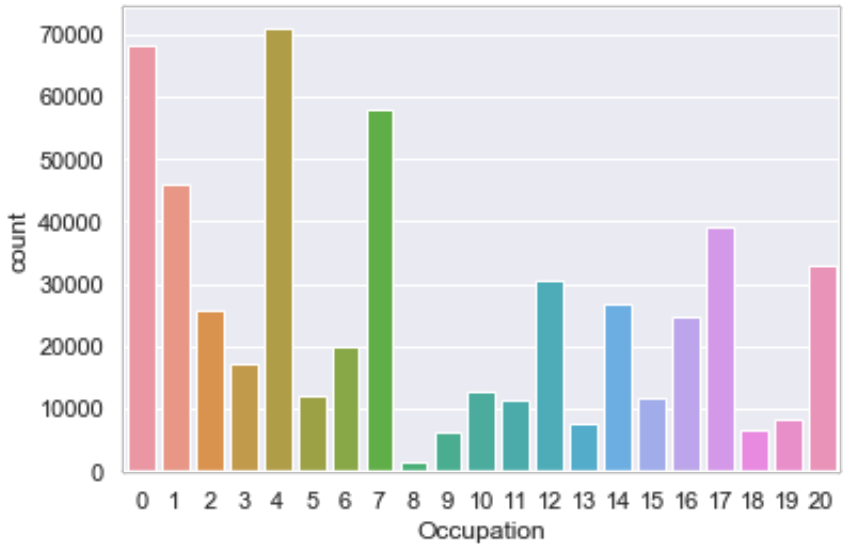
In [117]:
```
sns.countplot(ap['City_Category'])
```

Out[117]: <matplotlib.axes._subplots.AxesSubplot at 0x1a322c1160>



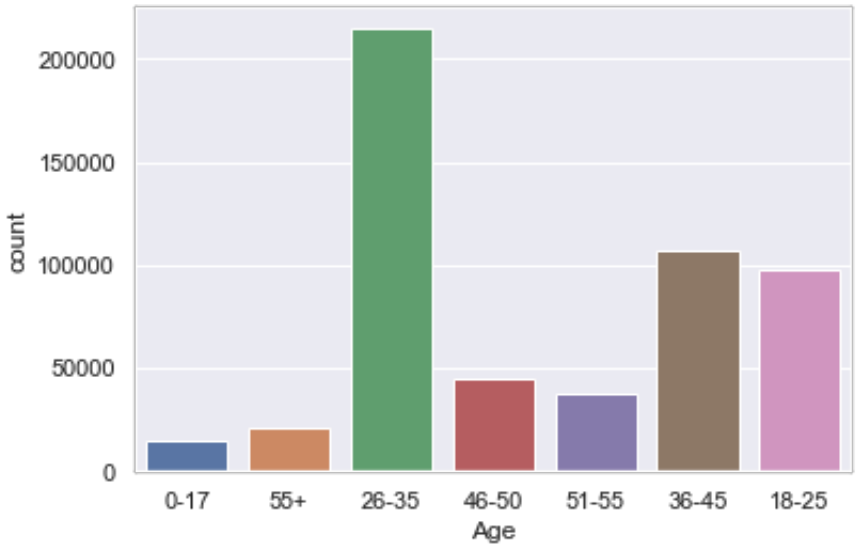
In [118]:
```
sns.countplot(ap['Occupation'])
```

Out[118]: <matplotlib.axes._subplots.AxesSubplot at 0x1a1a5bc320>



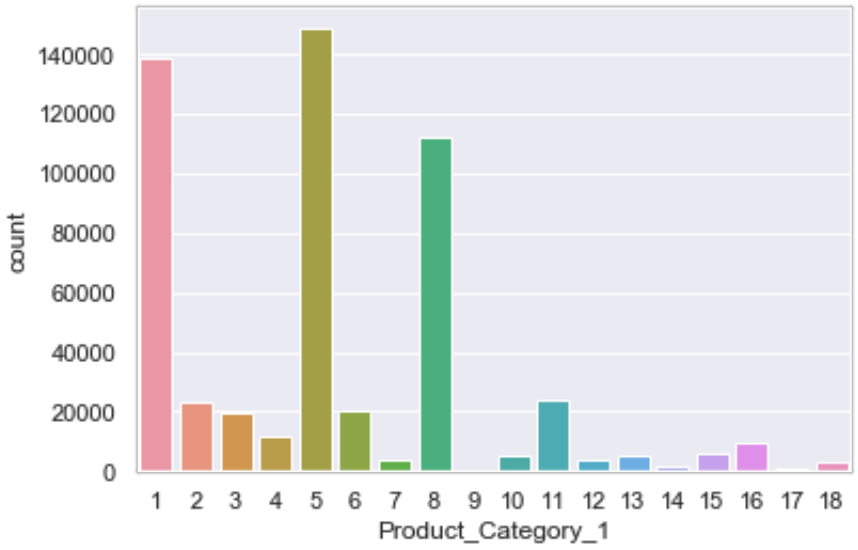
In [119]:
```
sns.countplot(ap['Age'])
```

Out[119]: <matplotlib.axes._subplots.AxesSubplot at 0x1a3235bfd0>



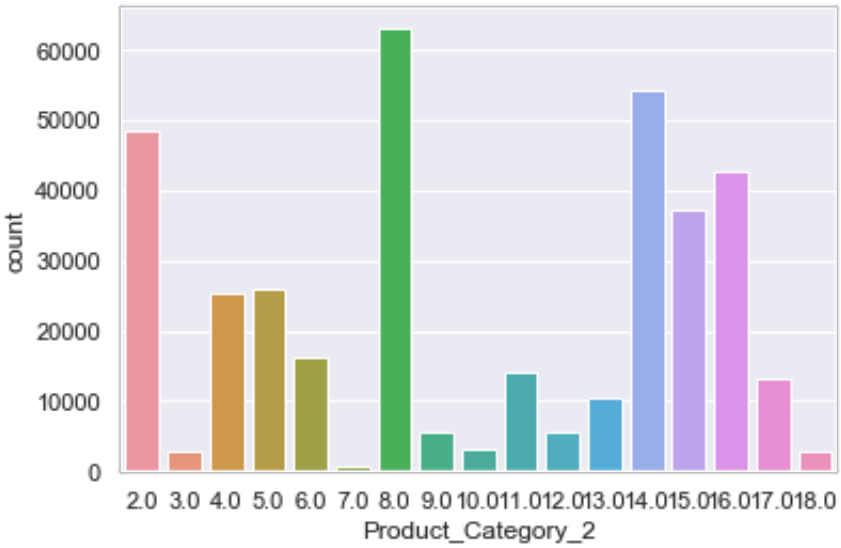
In [120]:
```
sns.countplot(ap['Product_Category_1'])
```

Out[120]: <matplotlib.axes._subplots.AxesSubplot at 0x1a32240e48>



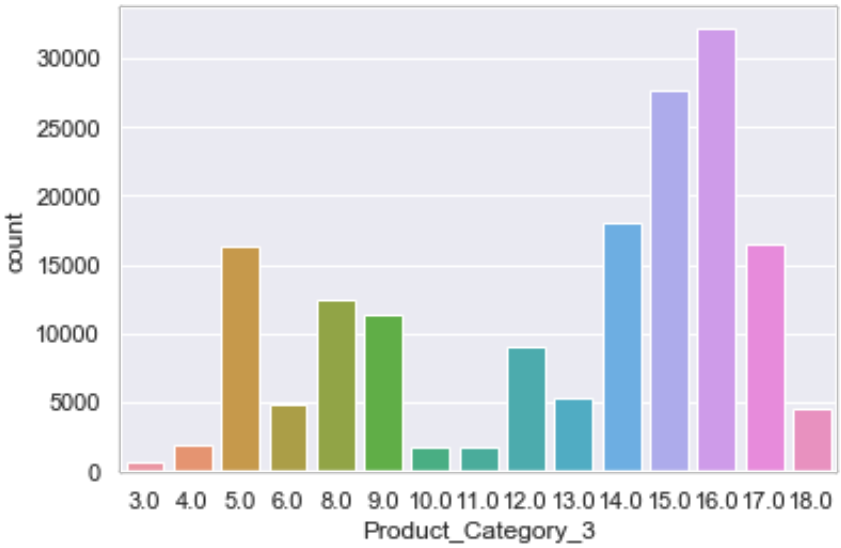
In [121]:
```
sns.countplot(ap['Product_Category_2'])
```

Out[121]: <matplotlib.axes._subplots.AxesSubplot at 0x1a2f5aa588>

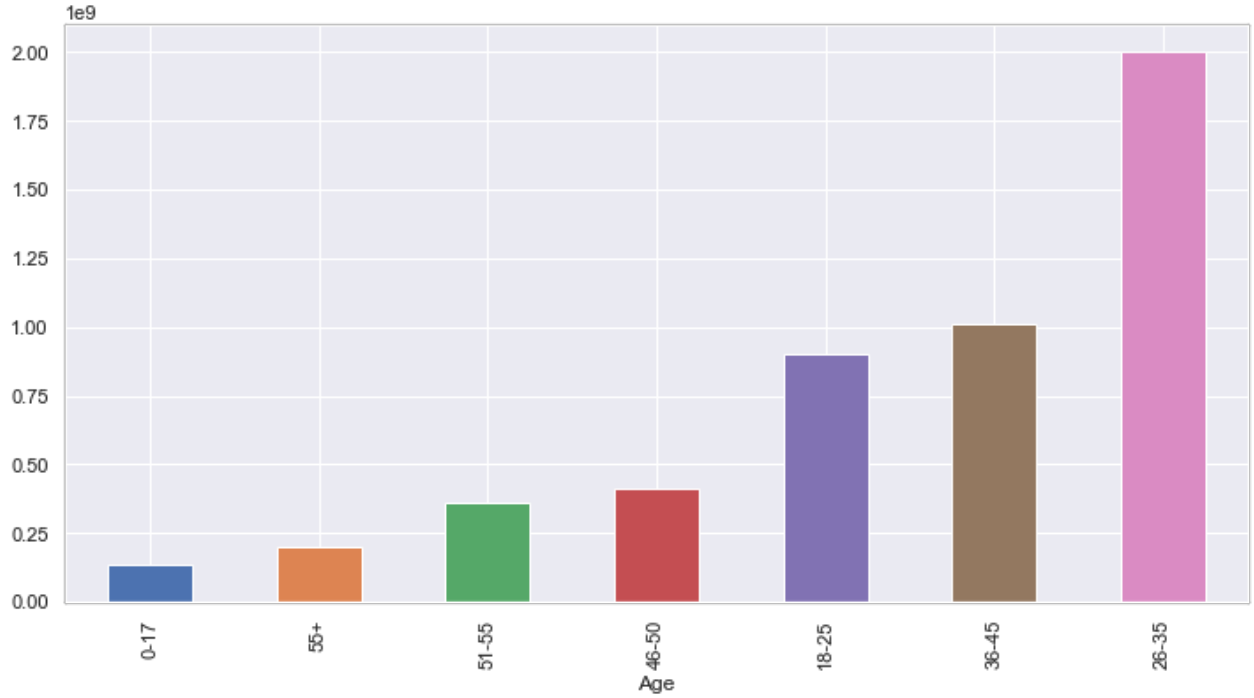


In [122]:
```
sns.countplot(ap['Product_Category_3'])
```

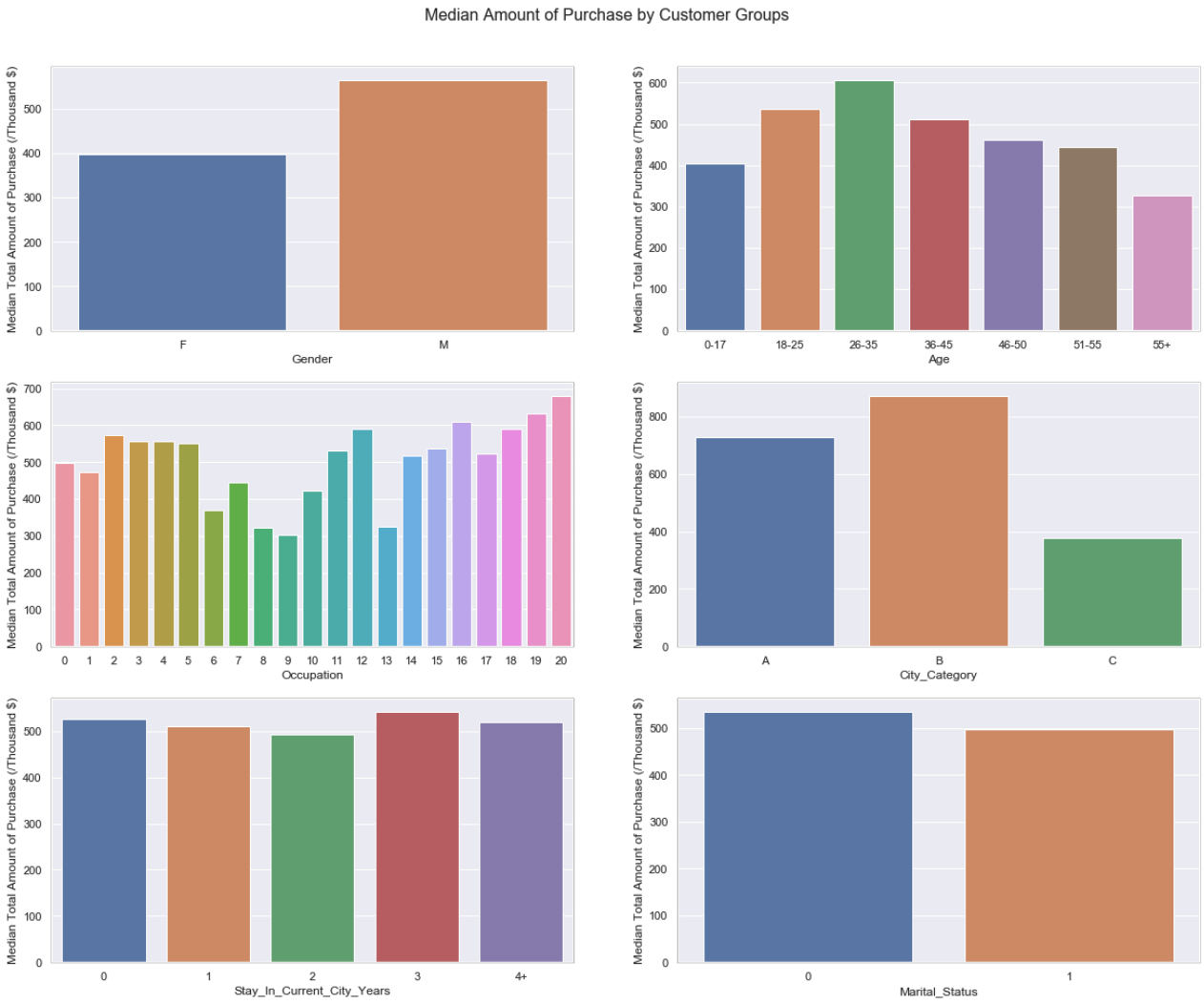
Out[122]: <matplotlib.axes._subplots.AxesSubplot at 0x1a322f7668>



In [123]:
```
fig1, ax1 = plt.subplots(figsize=(12,7))
sns.countplot(ap['Age'],hue=ap['Gender'])

def plot(group,column,plot):
    ax=plt.figure(figsize=(12,6))
    ap.groupby(group)[column].sum().sort_values().plot(plot)
plot('Age','Purchase','bar')
```



In [124]:
```
# Bar charts - show median instead of mean of total amount of purchase b
import numpy as np
fig5, axes = plt.subplots(3,2,figsize=(20,16))

fig5.suptitle('Median Amount of Purchase by Customer Groups', fontsize =

sns.barplot(x='Gender', y='Tot_Purchase', data = ap_customer, estimator =
sns.barplot(x='Age', y='Tot_Purchase', data = ap_customer, estimator = n
            ax = axes[0][1], order = ['0-17', '18-25', '26-35', '36-45',
sns.barplot(x='Occupation', y='Tot_Purchase', data = ap_customer, estima
sns.barplot(x='City_Category', y='Tot_Purchase', data = ap_customer, est
            ci = None, ax = axes[1][1], order = ('A', 'B', 'C'))
sns.barplot(x='Stay_In_Current_City_Years', y='Tot_Purchase', data = ap_
            ci = None, ax = axes[2][0], order = ('0', '1', '2', '3', '4+
sns.barplot(x='Marital_Status', y='Tot_Purchase', data = ap_customer, es

for ax in fig5.axes:
    plt.sca(ax)
    plt.ylabel('Median Total Amount of Purchase (/Thousand $)')

plt.savefig('fig5')
```

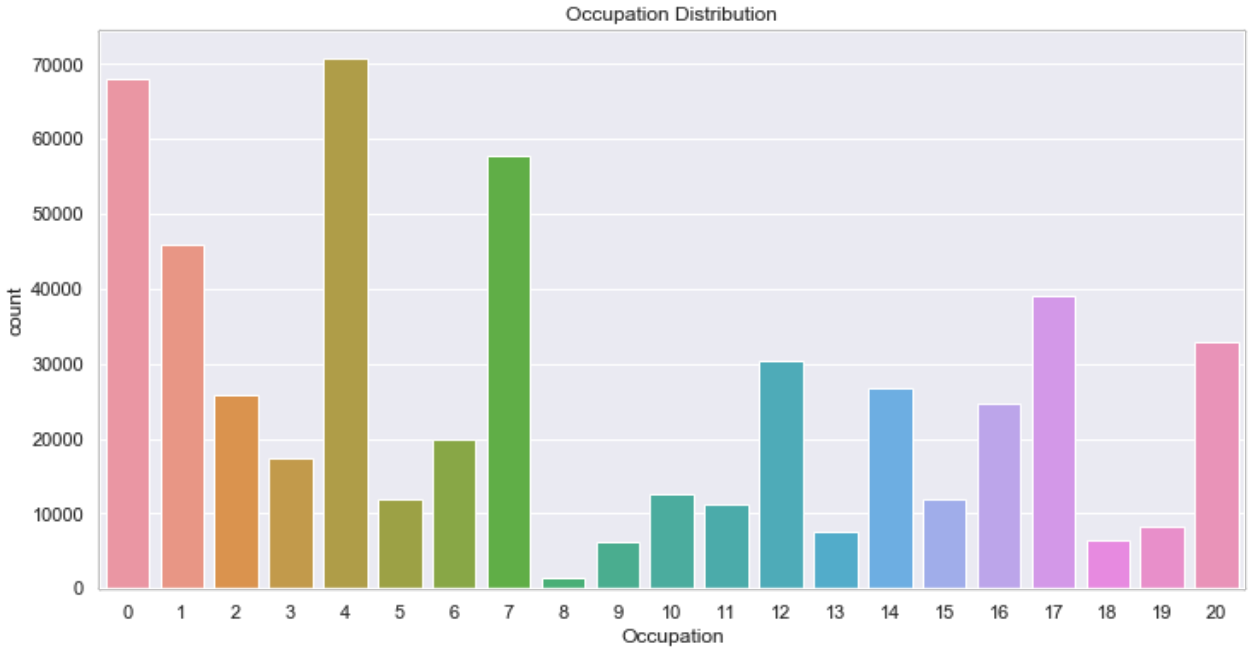

Median Amount of Purchase by Customer Groups
Median Total Amount of Purchase (/Thousand $)
Gender
F
M
Age
0-17
18-25
26-35
36-45
46-50
51-55
55+
Occupation
City_Category
A
B
C
Stay_In_Current_City_Years
Marital_Status

In [143]:
```
# Generate new features - total # products purchased by customer; and to
tot_item = ap['User_ID'].value_counts().sort_index()
tot_purchase = ap.groupby('User_ID').sum()['Purchase']
tot = pd.concat([tot_item, tot_purchase], axis = 1, keys = ['Tot_Product

ap = pd.merge(ap, tot, left_on = 'User_ID', right_index = True)
ap.head()
```

Out[143]:

...	Tot_Products_x	Tot_Purchase_x	Tot_Products_y	Tot_Purchase_y	Tot_Products_x	Tot_Purchase_
...	34	333481	34	333481	34	33348
...	34	333481	34	333481	34	33348
...	34	333481	34	333481	34	33348
...	34	333481	34	333481	34	33348
...	34	333481	34	333481	34	33348

In [144]:
```
#Occupation
plt.figure(figsize=(12,6))
sns.countplot(ap['Occupation'])
plt.title('Occupation Distribution')
plt.show()
```



Question :- 1) We need to Analysis the data based on the multiple variables on which the purchase is dependent. So we will analyze all variable such as Gender , Age, Occupation, City in which they Stay, Product Category from which they shop, Marital Status. 2) Once we study these small variable we will come to know the impact of each Variable on the purchase Level. Now from this we need to analyze on each user to check what was their total purchase. As this will give us a view of their total purchasing power. Now Analzing each variable again we will come to know which of them impact the most to affect the purchasing power of each Individual.

Question What we Exactly want to Predict :- 1) We want to check the purchasing power of user once they login with their Details. 2) We want to predict the customer recommendation based on product Category for each new user once we have their details. 3) Inbound goods prediction based on sale for each city category.

In [ ]: