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

In [31]:
```
ap = pd.read_csv("F:/Semester 2nd/Multi Variate/Project/Blackfriday.csv")
ap.head(10)
```

Out[31]:

	User_ID	Product_ID	Gender	Age	Occupation	City_Category	Stay_In_Current_City_Years	M
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 [32]:
```
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 [33]:
```
ap.isnull().sum()
```

Out[33]:
```
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 [35]:
```
ap.columns
```

Out[35]:
```
Index(['User_ID', 'Product_ID', 'Gender', 'Age', 'Occupation', 'City_Categor
y',
       'Stay_In_Current_City_Years', 'Marital_Status', 'Product_Category_1',
       'Product_Category_2', 'Product_Category_3', 'Purchase'],
      dtype='object')
```

In [36]:
```
ap.sort_values('User_ID').head(10)
ap['User_ID'].value_counts().count()
```

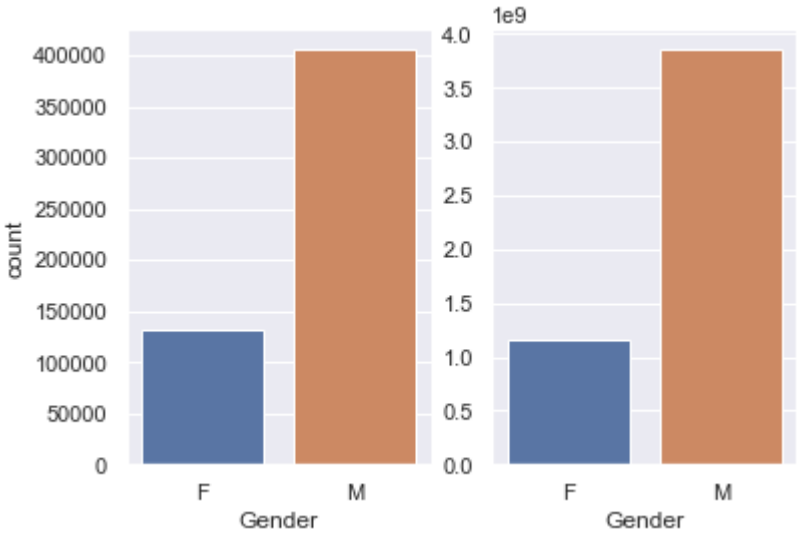
Out[36]:
```
5891
```

In [41]:
```
ap['Gender'].unique()
```

Out[41]:
```
array(['F', 'M'], dtype=object)
```

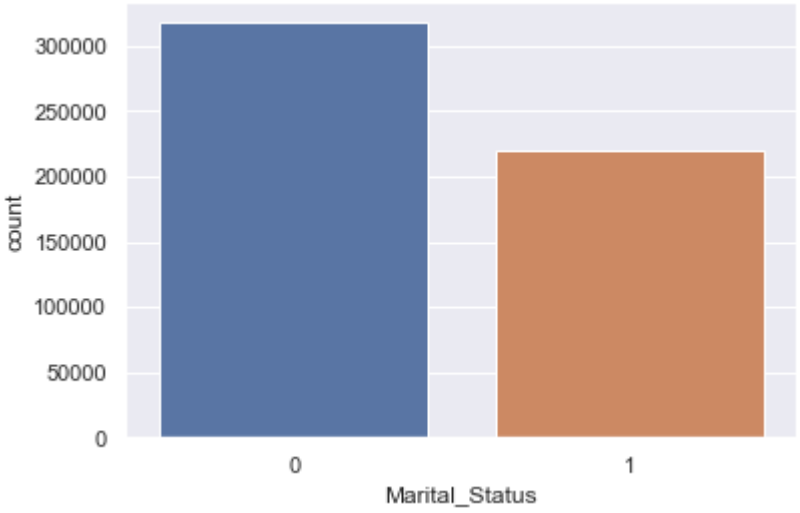
In [42]:
```
plt.subplot(1,2,1)
sns.countplot(ap['Gender']) #attendance

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

Out[42]: <matplotlib.axes._subplots.AxesSubplot at 0x1a16903630>



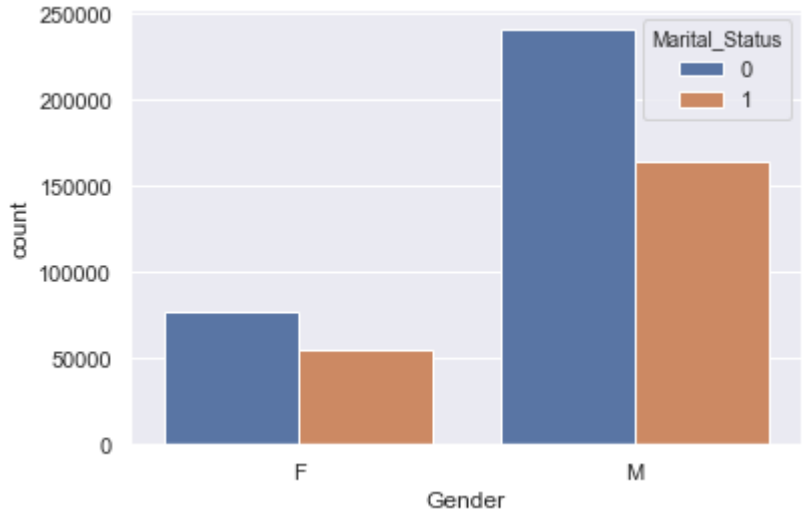
In [43]:
```
sns.countplot(ap['Marital_Status'])
```

Out[43]: <matplotlib.axes._subplots.AxesSubplot at 0x1a178d9d30>



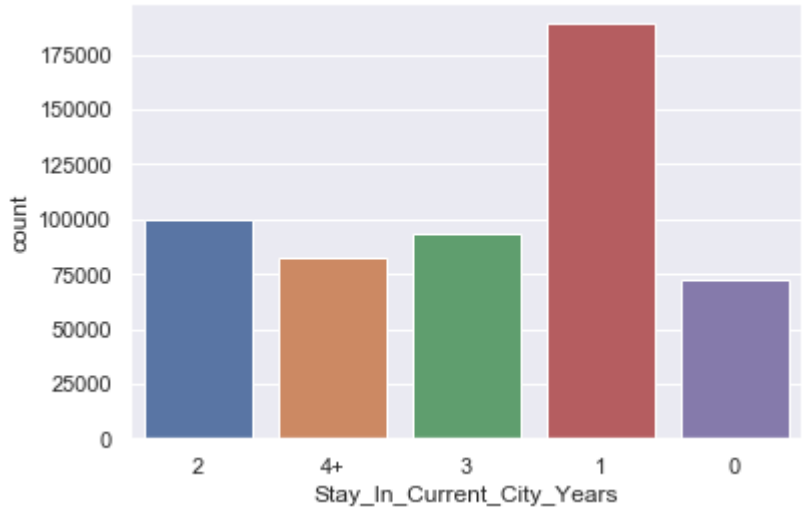
In [44]:
```
sns.countplot(ap['Gender'], hue = ap['Marital_Status'])
```

Out[44]: <matplotlib.axes._subplots.AxesSubplot at 0x1a17fedcf8>



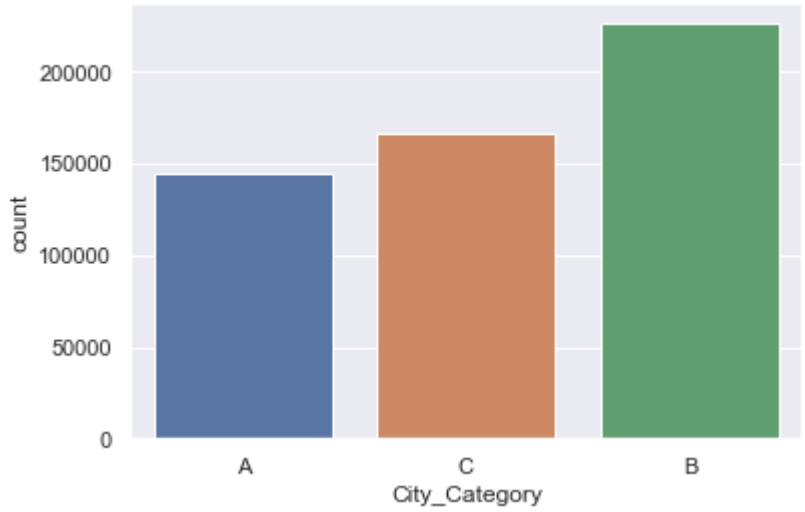
In [45]:
```
sns.countplot(ap['Stay_In_Current_City_Years'])
```

Out[45]: <matplotlib.axes._subplots.AxesSubplot at 0x1a18cb1b00>



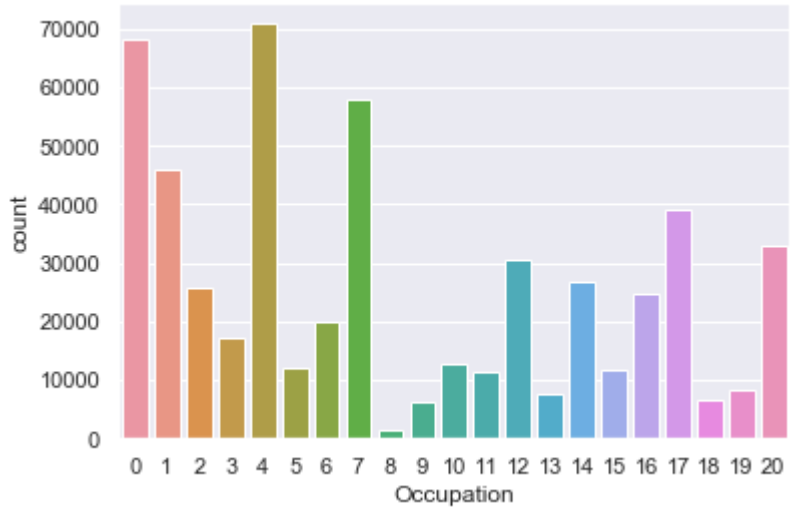
In [46]:
```
sns.countplot(ap['City_Category'])
```

Out[46]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1a192c0860>
```



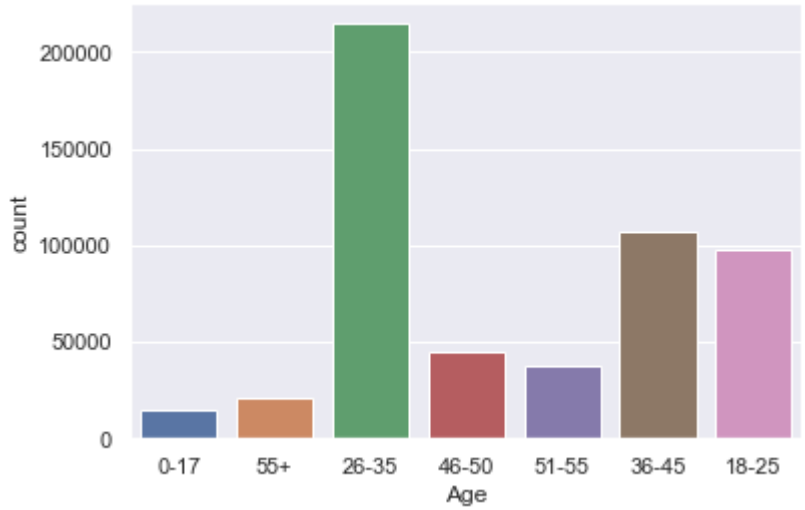
In [47]:
```
sns.countplot(ap['Occupation'])
```

Out[47]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1a188da828>
```



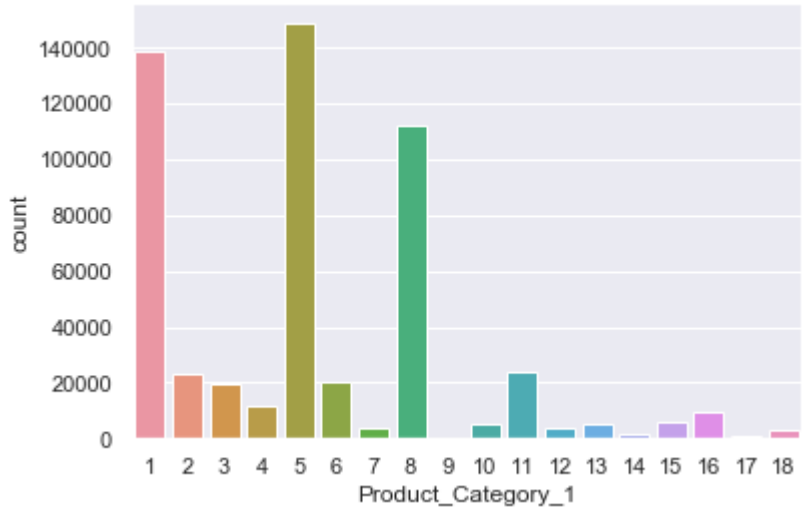
In [48]:
```
sns.countplot(ap['Age'])
```

Out[48]: <matplotlib.axes._subplots.AxesSubplot at 0x1a1a75c470>



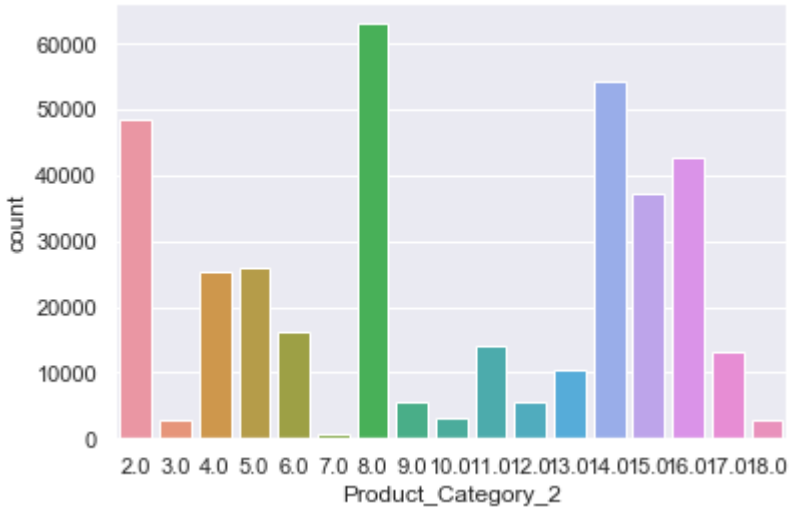
In [49]:
```
sns.countplot(ap['Product_Category_1'])
```

Out[49]: <matplotlib.axes._subplots.AxesSubplot at 0x1a19fd3dd8>



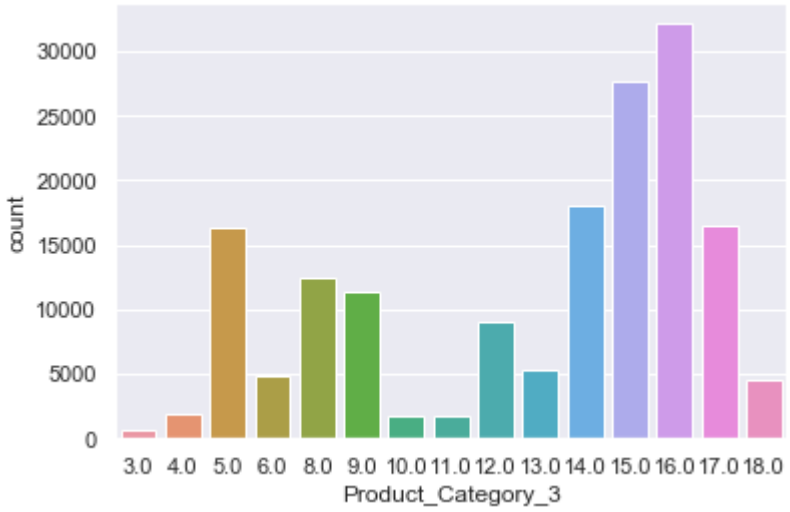
In [50]:
```
sns.countplot(ap['Product_Category_2'])
```

Out[50]: <matplotlib.axes._subplots.AxesSubplot at 0x1a1924f358>

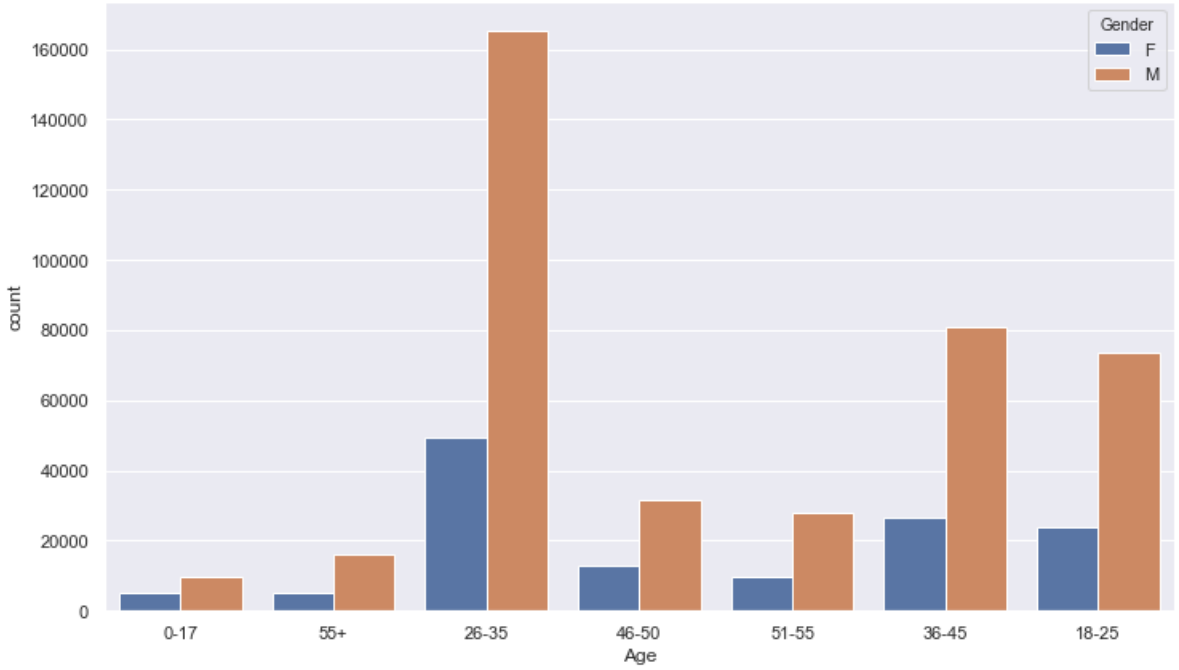


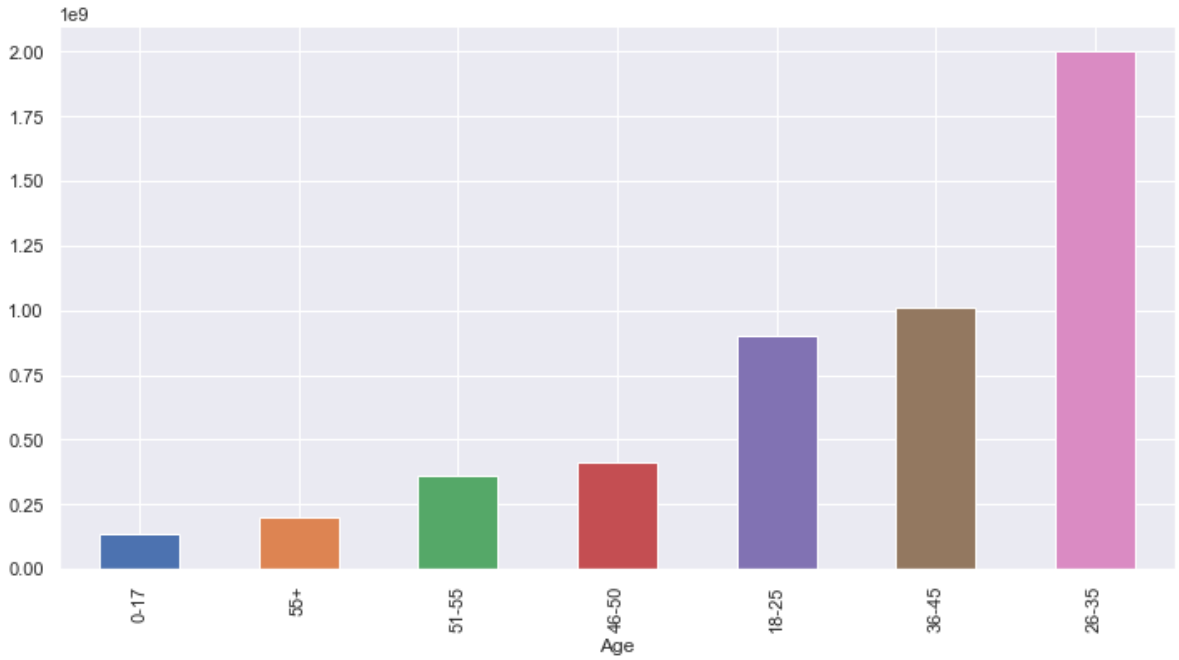
In [51]:
```
sns.countplot(ap['Product_Category_3'])
```

Out[51]: <matplotlib.axes._subplots.AxesSubplot at 0x1a1920b668>

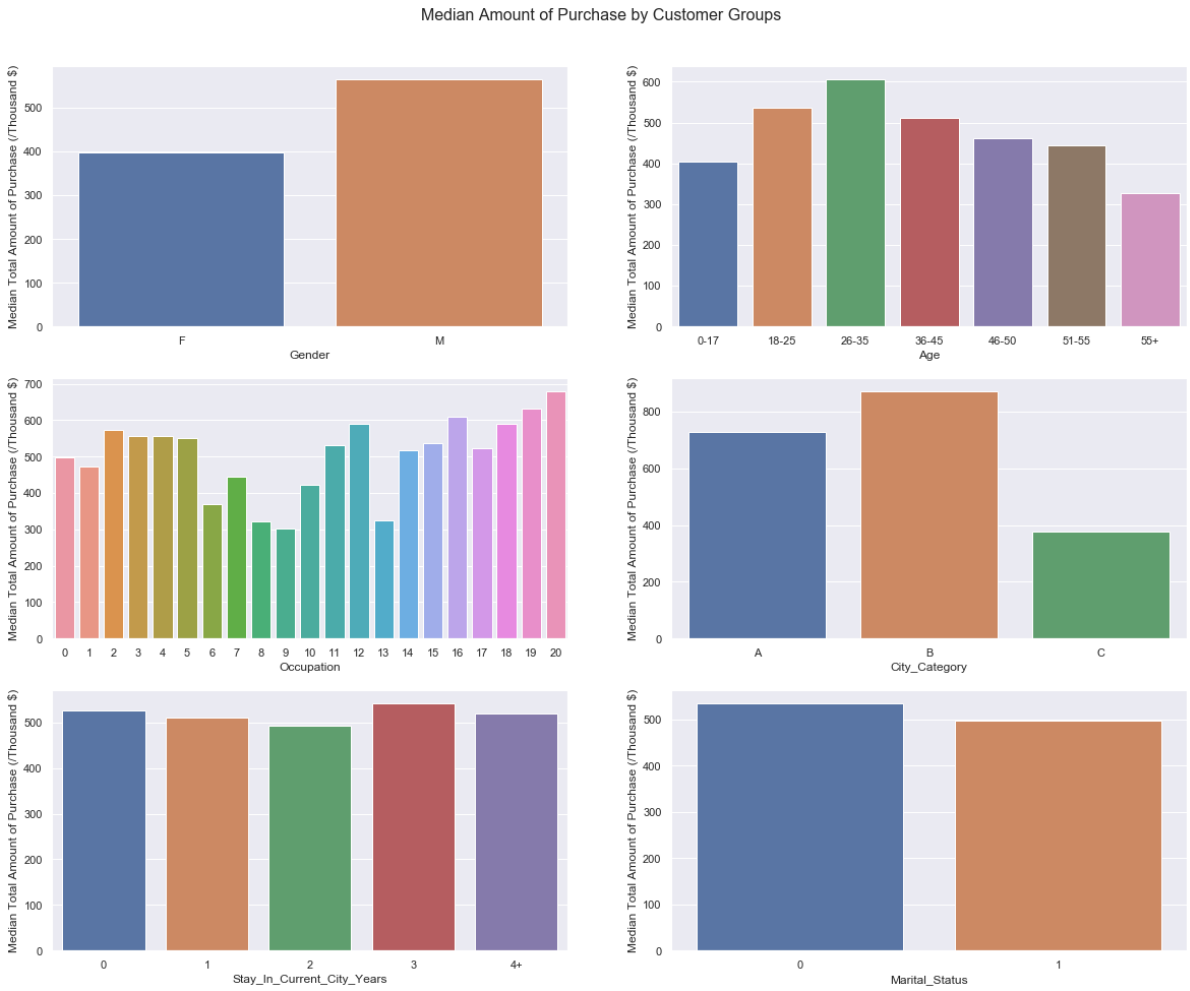


In [52]:
```
fig1, ax1 = plt.subplots(figsize=(12,7))
sns.countplot(ap['Age'],hue=ap['Gender'])

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 [70]:
```
# Bar charts - show median instead of mean of total amount of purchase by each
characteristic
import numpy as np
fig5, axes = plt.subplots(3,2,figsize=(20,16))

fig5.suptitle('Median Amount of Purchase by Customer Groups', fontsize = 16, y
= 0.93)

sns.barplot(x='Gender', y='Tot_Purchase', data = ap_customer, estimator = np.m
edian, ci = None, ax = axes[0][0])
sns.barplot(x='Age', y='Tot_Purchase', data = ap_customer, estimator = np.medi
an, ci = None,
            ax = axes[0][1], order = ['0-17', '18-25', '26-35', '36-45', '46-5
0', '51-55', '55+'])
sns.barplot(x='Occupation', y='Tot_Purchase', data = ap_customer, estimator =
np.median, ci = None, ax = axes[1][0])
sns.barplot(x='City_Category', y='Tot_Purchase', data = ap_customer, estimator
= np.median,
            ci = None, ax = axes[1][1], order = ('A', 'B', 'C'))
sns.barplot(x='Stay_In_Current_City_Years', y='Tot_Purchase', data = ap_custom
er, estimator = np.median,
            ci = None, ax = axes[2][0], order = ('0', '1', '2', '3', '4+'))
sns.barplot(x='Marital_Status', y='Tot_Purchase', data = ap_customer, estimato
r = np.median, ci = None, ax = axes[2][1])

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 [ ]: