In [1]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
```

In [2]:

```
import os
os.getcwd()
```

Out[2]:

```
'C:\\Users\\ADHYATM MISHRA\\Data-Science with python'
```

In [3]:

```
os.chdir("C:\\Users\\ADHYATM MISHRA")
```

In [4]:

```
os.getcwd()
```

Out[4]:

```
'C:\\Users\\ADHYATM MISHRA'
```

In [5]:

```
df=pd.read_csv('diwali.csv',encoding= 'unicode_escape')
```

In [6]:

```
df.head()
```

Out[6]:

	User_ID	Cust_name	Product_ID	Gender	Age Group	Age	Marital_Status	State	Zone	Occupation	Product_Categ
0	1002903	Sanskriti	P00125942	F	26-35	28	0	Maharashtra	Western	Healthcare	A
1	1000732	Kartik	P00110942	F	26-35	35	1	Andhra Pradesh	Southern	Govt	A
2	1001990	Bindu	P00118542	F	26-35	35	1	Uttar Pradesh	Central	Automobile	A
3	1001425	Sudevi	P00237842	M	0-17	16	0	Karnataka	Southern	Construction	A
4	1000588	Joni	P00057942	M	26-35	28	1	Gujarat	Western	Food Processing	A

In [7]:

```
df.shape
```

Out[7]:

```
(11251, 15)
```

In [8]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11251 entries, 0 to 11250
Data columns (total 15 columns):
User_ID             11251 non-null int64
Cust_name           11251 non-null object
Product_ID          11251 non-null object
Gender              11251 non-null object
Age Group           11251 non-null object
Age                 11251 non-null int64
Marital_Status      11251 non-null int64
State               11251 non-null object
Zone                11251 non-null object
Occupation          11251 non-null object
Product_Category    11251 non-null object
Orders              11251 non-null int64
Amount              11239 non-null float64
Status              0 non-null float64
unnamed1            0 non-null float64
dtypes: float64(3), int64(4), object(8)
memory usage: 1.3+ MB
```

In [9]:

```
df.isnull().sum()
```

Out[9]:

```
User_ID                 0
Cust_name               0
Product_ID              0
Gender                  0
Age Group               0
Age                     0
Marital_Status          0
State                   0
Zone                    0
Occupation              0
Product_Category        0
Orders                  0
Amount                 12
Status              11251
unnamed1            11251
dtype: int64
```

In [10]:

```
df.drop(['Status','unnamed1'],axis =1,inplace=True)
```

In [11]:

```
df.head(2)
```

Out[11]:

	User_ID	Cust_name	Product_ID	Gender	Age Group	Age	Marital_Status	State	Zone	Occupation	Product_Categ
0	1002903	Sanskriti	P00125942	F	26-35	28	0	Maharashtra	Western	Healthcare	A
1	1000732	Kartik	P00110942	F	26-35	35	1	Andhra Pradesh	Southern	Govt	A

In [12]:

```
df.dropna(inplace=True)
```

In [13]:

```
df.isnull().sum()
```

Out[13]:

```
User_ID             0
Cust_name           0
Product_ID          0
Gender              0
Age Group           0
Age                 0
Marital_Status      0
State               0
Zone                0
Occupation          0
Product_Category    0
Orders              0
Amount              0
dtype: int64
```

In [14]:

```
df.shape
```

Out[14]:

```
(11239, 13)
```

In [15]:

```
df['Amount']=df['Amount'].astype('int')
```

In [16]:

```
df['Amount'].dtypes
```

Out[16]:

```
dtype('int32')
```

In [17]:

```
df.columns
```

Out[17]:

```
Index(['User_ID', 'Cust_name', 'Product_ID', 'Gender', 'Age Group', 'Age',
       'Marital_Status', 'State', 'Zone', 'Occupation', 'Product_Category',
       'Orders', 'Amount'],
      dtype='object')
```

In [18]:

```
df[['Age','Amount','Orders']].describe()
```

Out[18]:

	Age	Amount	Orders
count	11239.000000	11239.000000	11239.000000
mean	35.410357	9453.610553	2.489634
std	12.753866	5222.355168	1.114967
min	12.000000	188.000000	1.000000
25%	27.000000	5443.000000	2.000000
50%	33.000000	8109.000000	2.000000
75%	43.000000	12675.000000	3.000000
max	92.000000	23952.000000	4.000000

Exploratory Data Analysis

Gender

In [19]:

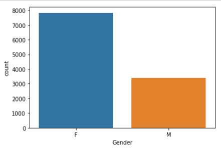
```
df.columns
```

Out[19]:

```
Index(['User_ID', 'Cust_name', 'Product_ID', 'Gender', 'Age Group', 'Age',
       'Marital_Status', 'State', 'Zone', 'Occupation', 'Product_Category',
       'Orders', 'Amount'],
      dtype='object')
```

In [20]:

```
# Count of Male V/S Female Buyers
ax= sns.countplot(x=df.Gender,data=df)
```



In [21]:

```
# Amount spent Male V/S Female Buyers
sales_gen=df.groupby(['Gender'],as_index=False)['Amount'].sum().sort_values(by='Amount',asce
nding=False)
sales_gen
```

Out[21]:

	Gender	Amount
0	F	74335853
1	M	31913276

From above graphs we can see that most of the buyers are females and even the purchasing power of females are greater than men.

Age

In [22]:

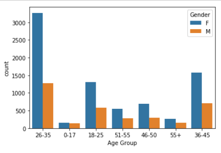
```
df.columns
```

Out[22]:

```
Index(['User_ID', 'Cust_name', 'Product_ID', 'Gender', 'Age Group', 'Age',
       'Marital_Status', 'State', 'Zone', 'Occupation', 'Product_Category',
       'Orders', 'Amount'],
      dtype='object')
```

In [23]:

```
# Count of Purchasing power by Age group
ax = sns.countplot(data = df, x = 'Age Group', hue = 'Gender')
```



In [24]:

```
# Total Amount vs Age Group in table format
sales_gen=df.groupby(['Age Group'],as_index=False)['Amount'].sum().sort_values(by='Amount',a
scending=False)
sales_gen
```

Out[24]:

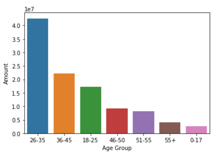
	Age Group	Amount
2	26-35	42613442
3	36-45	22144994
1	18-25	17240732
4	46-50	9207844
5	51-55	8261477
6	55+	4080987
0	0-17	2699653

In [25]:

```
# Total Amount vs Age Group in graph format
sns.barplot(x='Age Group',y='Amount',data=sales_gen)
```

Out[25]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1ee056c0888>
```



From above graphs we can see that most of the buyers are of age group between 26-35 yrs female

State

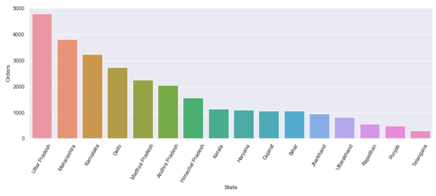
In [26]:

```
# total no. of orders from top 10 states

sales_state = df.groupby(['State'],as_index=False)['Orders'].sum().sort_values(by='Orders',a
scending=False)
sns.set(rc={'figure.figsize':(15,5)})
sns.barplot(data=sales_state,x='State',y='Orders')
plt.xticks(rotation=60)
```

Out[26]:

```
(array([ 0,  1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11, 12, 13, 14, 15]),
 <a list of 16 Text xticklabel objects>)
```



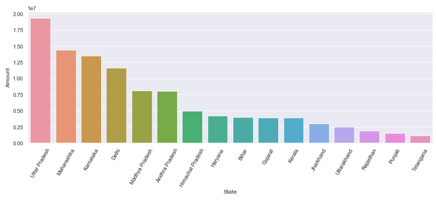
In [27]:

```
# total no. of Amount spent from top 10 states

sales_state = df.groupby(['State'],as_index=False)['Amount'].sum().sort_values(by='Amount',a
scending=False)
sns.set(rc={'figure.figsize':(15,5)})
sns.barplot(data=sales_state,x='State',y='Amount')
plt.xticks(rotation=60)
```

Out[27]:

```
(array([ 0,  1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11, 12, 13, 14, 15]),
 <a list of 16 Text xticklabel objects>)
```

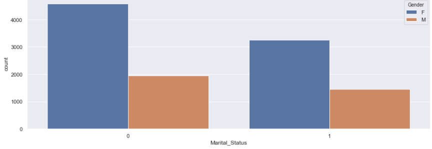


From above graphs we can see that most of the orders & total sales/amount are from Uttar Pradesh, Maharashtra and Karnataka respectively

Marital Status

In [28]:

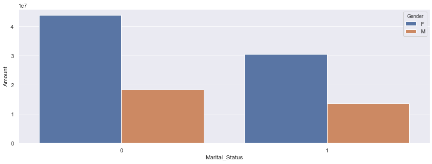
```
# countplot acc to Marital status and Gender
ax = sns.countplot(data=df,x=df.Marital_Status,hue='Gender')
```



In [29]:

```
# Countplot acc Amount spent by Marital status and Gender
Marital_stat = df.groupby(['Marital_Status','Gender'],as_index=False)['Amount'].sum().sort_v
alues(by='Amount',ascending=False)
sns.barplot(data=Marital_stat,x='Marital_Status',y='Amount',hue='Gender')
```

```
Out[29]: <matplotlib.axes._subplots.AxesSubplot at 0x1ee0596e148>
```

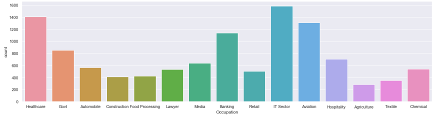


From above graphs we can see that most of the buyers are married (women) and they have high purchasing power

Occupation

In [30]:

```
# Countplot acc no. of time shoping done by occupation
sns.set(rc={'figure.figsize':(20,5)})
ax = sns.countplot(data = df, x = 'Occupation')
```

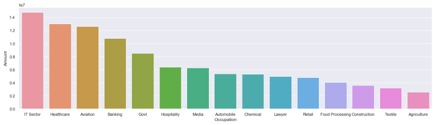


In [31]:

```
# Occupation by amount spent
occupation_stat = df.groupby(['Occupation'], as_index=False)['Amount'].sum().sort_values(by
='Amount', ascending=False)

sns.set(rc={'figure.figsize':(20,5)})
sns.barplot(data = occupation_stat, x = 'Occupation',y= 'Amount')
```

```
Out[31]: <matplotlib.axes._subplots.AxesSubplot at 0x1ee05ea9788>
```



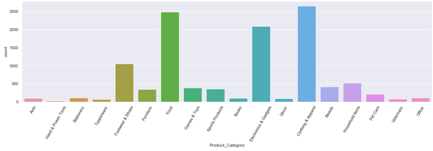
From above graphs we can see that most of the buyers are working in IT, Healthcare and Aviation sector

Product Category

In [32]:

```
# Count of products
sns.set(rc={'figure.figsize':(20,5)})
ax = sns.countplot(data = df, x = 'Product_Category')
plt.xticks(rotation=60)
```

```
Out[32]: (array([ 0,  1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11, 12, 13, 14, 15, 16,
        17]), <a list of 18 Text xticklabel objects>)
```

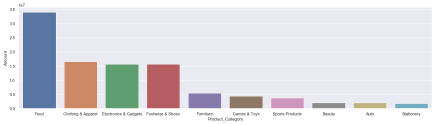


In [33]:

```
# Top sold product categories
Product_stat = df.groupby(['Product_Category'], as_index=False)['Amount'].sum().sort_values
(by='Amount', ascending=False).head(10)

sns.set(rc={'figure.figsize':(20,5)})
sns.barplot(data = Product_stat, x = 'Product_Category',y= 'Amount')
```

```
Out[33]: <matplotlib.axes._subplots.AxesSubplot at 0x1ee06165808>
```



From above graphs we can see that most of the sold products are from Food, Clothing and Electronics category

Conclusion:

Married women age group 26-35 yrs from UP, Maharastra and Karnataka working in IT, Healthcare and Aviation are more likely to buy products from Food, Clothing and Electronics category.