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

data=pd.read_csv("C:\Users\Dell\Desktop\Diwali_Sales_Data.csv",encoding= 'unicode_escape')
data
```

Out[1]:

	User_ID	Cust_name	Product_ID	Gender	Age Group	Age	Marital_Status	State	Zone	Occupation	Product_Category	Orders	Amount	Status	unnamed1
0	1002903	Sanskriti	P00125942	F	26-35	28	0	Maharashtra	Western	Healthcare	Auto	1	23952.0	NaN	NaN
1	1000732	Kartik	P00110942	F	26-35	35	1	Andhra Pradesh	Southern	Govt	Auto	3	23934.0	NaN	NaN
2	1001990	Bindu	P00110942	F	26-35	35	1	Uttar Pradesh	Central	Automobile	Auto	3	23924.0	NaN	NaN
3	1001425	Sudevi	P00237842	M	0-17	16	0	Karnataka	Southern	Construction	Auto	2	23912.0	NaN	NaN
4	1000588	Joni	P00057942	M	26-35	28	1	Gujarat	Western	Food Processing	Auto	2	23877.0	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
11246	1000695	Manning	P00296942	M	18-25	19	1	Maharashtra	Western	Chemical	Office	4	370.0	NaN	NaN
11247	1004089	Reichenbach	P00171342	M	26-35	33	0	Haryana	Northern	Healthcare	Veterinary	3	367.0	NaN	NaN
11248	1001209	Oshin	P00125042	F	36-45	40	0	Madhya Pradesh	Central	Textile	Office	4	213.0	NaN	NaN
11249	1004023	Noonan	P00095942	M	36-45	37	0	Karnataka	Southern	Agriculture	Office	3	206.0	NaN	NaN
11250	1002744	Brumley	P00255842	M	18-25	19	0	Maharashtra	Western	Healthcare	Office	3	188.0	NaN	NaN

11251 rows × 15 columns

In [2]:
```
data.shape
```

Out[2]:
(11251, 15)

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

Out[3]:

	User_ID	Cust_name	Product_ID	Gender	Age Group	Age	Marital_Status	State	Zone	Occupation	Product_Category	Orders	Amount	Status	unnamed1
0	1002903	Sanskriti	P00125942	F	26-35	28	0	Maharashtra	Western	Healthcare	Auto	1	23952.0	NaN	NaN
1	1000732	Kartik	P00110942	F	26-35	35	1	Andhra Pradesh	Southern	Govt	Auto	3	23934.0	NaN	NaN
2	1001990	Bindu	P00110942	F	26-35	35	1	Uttar Pradesh	Central	Automobile	Auto	3	23924.0	NaN	NaN
3	1001425	Sudevi	P00237842	M	0-17	16	0	Karnataka	Southern	Construction	Auto	2	23912.0	NaN	NaN
4	1000588	Joni	P00057942	M	26-35	28	1	Gujarat	Western	Food Processing	Auto	2	23877.0	NaN	NaN

In [4]:
```
data.head(12)
```

Out[4]:

	User_ID	Cust_name	Product_ID	Gender	Age Group	Age	Marital_Status	State	Zone	Occupation	Product_Category	Orders	Amount	Status	unnamed1
0	1002903	Sanskriti	P00125942	F	26-35	28	0	Maharashtra	Western	Healthcare	Auto	1	23952.0	NaN	NaN
1	1000732	Kartik	P00110942	F	26-35	35	1	Andhra Pradesh	Southern	Govt	Auto	3	23934.0	NaN	NaN
2	1001990	Bindu	P00110942	F	26-35	35	1	Uttar Pradesh	Central	Automobile	Auto	3	23924.0	NaN	NaN
3	1001425	Sudevi	P00237842	M	0-17	16	0	Karnataka	Southern	Construction	Auto	2	23912.0	NaN	NaN
4	1000588	Joni	P00057942	M	26-35	28	1	Gujarat	Western	Food Processing	Auto	2	23877.0	NaN	NaN
5	1000588	Joni	P00057942	M	26-35	28	1	Himachal Pradesh	Northern	Food Processing	Auto	1	23877.0	NaN	NaN
6	1001132	Balk	P00018042	F	18-25	25	1	Uttar Pradesh	Central	Lawyer	Auto	4	23841.0	NaN	NaN
7	1002092	Shivangi	P00372442	F	55+	61	0	Maharashtra	Western	IT Sector	Auto	1	NaN	NaN	NaN
8	1002224	Kare	P00255942	F	26-35	35	0	Uttar Pradesh	Central	Govt	Auto	2	23836.0	NaN	NaN
9	1003650	Ginny	P00031142	F	26-35	26	1	Andhra Pradesh	Southern	Media	Auto	4	23798.00	NaN	NaN
10	1003029	Harsha	P00255842	M	26-35	34	0	Delhi	Central	Banking	Auto	1	23770.00	NaN	NaN
11	1000214	Kargelis	P00118142	F	18-25	20	0	Andhra Pradesh	Southern	Retail	Auto	2	23752.00	NaN	NaN

In [5]:
```
#data cleaning
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11251 entries, 0 to 11250
Data columns (total 15 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   User_ID         11251 non-null  int64
 1   Cust_name       11251 non-null  object
 2   Product_ID      11251 non-null  object
 3   Gender          11251 non-null  object
 4   Age Group       11251 non-null  object
 5   Age             11251 non-null  int64
 6   Marital_Status  11251 non-null  int64
 7   State           11251 non-null  object
 8   Zone            11251 non-null  object
 9   Occupation      11251 non-null  object
 10  Product_Category  11251 non-null  object
 11  Orders          11251 non-null  int64
 12  Amount          11239 non-null  float64
 13  Status          0 non-null      float64
 14  unnamed1        0 non-null      float64
dtypes: float64(3), int64(4), object(8)
memory usage: 1.3+ MB
```

In [6]:
```
# delete column
data.drop(['Status','unnamed1'], axis=1, inplace=True)
```

In [7]:
```
data.head(1)
```

Out[7]:

	User_ID	Cust_name	Product_ID	Gender	Age Group	Age	Marital_Status	State	Zone	Occupation	Product_Category	Orders	Amount
0	1002903	Sanskriti	P00125942	F	26-35	28	0	Maharashtra	Western	Healthcare	Auto	1	23952.0

In [8]:
```
#drop null values
data.dropna(inplace=True)
```

In [9]:
```
data.shape
```

Out[9]:
(11239, 13)

In [10]:
```
# drop null values in your columns
data.dropna(inplace=True)
```

In [11]:
```
data.shape
data.isnull().sum()
```

Out[11]:
```
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
dtype: int64
```

In [12]:
```
data.shape
```

Out[12]:
(11239, 13)

In [13]:
```
# drop null values in your columns
data.dropna(inplace=True)
```

In [14]:
```
data.shape
data.isnull().sum()
```

Out[14]:
```
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
```

Both are same

In [15]:
```
data.dropna(inplace = True) or
data1 = data.dropna()
```

In [16]:
```
# change data type
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 11239 entries, 0 to 11250
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   User_ID         11239 non-null  int64
 1   Cust_name       11239 non-null  object
 2   Product_ID      11239 non-null  object
 3   Gender          11239 non-null  object
 4   Age Group       11239 non-null  object
 5   Age             11239 non-null  int64
 6   Marital_Status  11239 non-null  int64
 7   State           11239 non-null  object
 8   Zone            11239 non-null  object
 9   Occupation      11239 non-null  object
 10  Product_Category  11239 non-null  object
 11  Orders          11239 non-null  int64
 12  Amount          11239 non-null  float64
dtypes: float64(1), int64(4), object(8)
memory usage: 1.2+ MB
```

In [17]:
```
# change data type on column 'Amount'
data['Amount']=data['Amount'].astype('int')
```

In [18]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 11239 entries, 0 to 11250
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   User_ID         11239 non-null  int64
 1   Cust_name       11239 non-null  object
 2   Product_ID      11239 non-null  object
 3   Gender          11239 non-null  object
 4   Age Group       11239 non-null  object
 5   Age             11239 non-null  int64
 6   Marital_Status  11239 non-null  int64
 7   State           11239 non-null  object
 8   Zone            11239 non-null  object
 9   Occupation      11239 non-null  object
 10  Product_Category  11239 non-null  object
 11  Orders          11239 non-null  int64
 12  Amount          11239 non-null  int32
dtypes: int32(1), int64(4), object(8)
memory usage: 1.2+ MB
```

In [19]:
```
data.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]:
```
data.rename(columns={'Marital_Status':'Status'}, inplace =True)
data
```

Out[20]:

	User_ID	Cust_name	Product_ID	Gender	Age Group	Age	Status	State	Zone	Occupation	Product_Category	Orders	Amount
0	1002903	Sanskriti	P00125942	F	26-35	28	0	Maharashtra	Western	Healthcare	Auto	1	23952
1	1000732	Kartik	P00110942	F	26-35	35	1	Andhra Pradesh	Southern	Govt	Auto	3	23934
2	1001990	Bindu	P00110942	F	26-35	35	1	Uttar Pradesh	Central	Automobile	Auto	3	23924
3	1001425	Sudevi	P00237842	M	0-17	16	0	Karnataka	Southern	Construction	Auto	2	23912
4	1000588	Joni	P00057942	M	26-35	28	1	Gujarat	Western	Food Processing	Auto	2	23877
...	...	...	...	...	...	...	...	...	...	...	...	...	...
11246	1000695	Manning	P00296942	M	18-25	19	1	Maharashtra	Western	Chemical	Office	4	370
11247	1004089	Reichenbach	P00171342	M	26-35	33	0	Haryana	Northern	Healthcare	Veterinary	3	367
11248	1001209	Oshin	P00125042	F	36-45	40	0	Madhya Pradesh	Central	Textile	Office	4	213
11249	1004023	Noonan	P00095942	M	36-45	37	0	Karnataka	Southern	Agriculture	Office	3	206
11250	1002744	Brumley	P00255842	M	18-25	19	0	Maharashtra	Western	Healthcare	Office	3	188

11239 rows × 13 columns

In [21]:
```
data.head()
```

Out[21]:

	User_ID	Cust_name	Product_ID	Gender	Age Group	Age	Status	State	Zone	Occupation	Product_Category	Orders	Amount
0	1002903	Sanskriti	P00125942	F	26-35	28	0	Maharashtra	Western	Healthcare	Auto	1	23952
1	1000732	Kartik	P00110942	F	26-35	35	1	Andhra Pradesh	Southern	Govt	Auto	3	23934
2	1001990	Bindu	P00110942	F	26-35	35	1	Uttar Pradesh	Central	Automobile	Auto	3	23924
3	1001425	Sudevi	P00237842	M	0-17	16	0	Karnataka	Southern	Construction	Auto	2	23912
4	1000588	Joni	P00057942	M	26-35	28	1	Gujarat	Western	Food Processing	Auto	2	23877

In [22]:
```
# data describe as all numeric values having some relavent info like mean,median,mode,max ,min,sd
data.describe()
```

Out[22]:

	User_ID	Age	Status	Orders	Amount
count	1.123900e+04	11239.000000	11239.000000	11239.000000	11239.000000
mean	1.003004e+06	35.410357	0.420055	2.489634	9453.610553
std	1.716039e+03	12.753866	0.493589	1.114782	5222.355168
min	1.000001e+06	12.000000	0.000000	1.000000	188.000000
25%	1.001407e+06	27.000000	0.000000	2.000000	5443.000000
50%	1.003064e+06	33.000000	0.000000	2.000000	8109.000000
75%	1.004428e+06	43.000000	1.000000	3.000000	12675.000000
max	1.006040e+06	92.000000	1.000000	4.000000	23952.000000

In [23]:
```
# particular column used describe
data['Age','Orders','Amount'].describe()
```

Out[23]:

	Age	Orders	Amount
count	11239.000000	11239.000000	11239.000000
mean	35.410357	2.489634	9453.610553
std	12.753866	1.114782	5222.355168
min	12.000000	1.000000	188.000000
25%	27.000000	2.000000	5443.000000
50%	33.000000	2.000000	8109.000000
75%	43.000000	3.000000	12675.000000
max	92.000000	4.000000	23952.000000

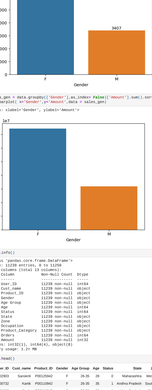
ETL

In [24]:
```
ax = sns.countplot(x = 'Gender',data= data)

# add labels

for bars in ax.containers:
    ax.bar_label(bars)
```

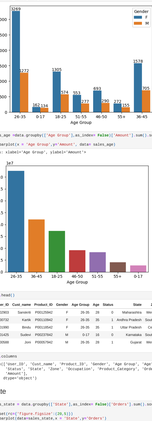
Out[24]:



In [25]:
```
sales_gen = data.groupby(['Gender'],as_index= False)['Amount'].sum().sort_values(by='Amount', ascending=False)
sns.barplot(x='Gender',y='Amount',data = sales_gen)

#bars.bar_label('Gender', ylabel='Amount')
```

Out[25]:



In [26]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 11239 entries, 0 to 11250
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   User_ID         11239 non-null  int64
 1   Cust_name       11239 non-null  object
 2   Product_ID      11239 non-null  object
 3   Gender          11239 non-null  object
 4   Age Group       11239 non-null  object
 5   Age             11239 non-null  int64
 6   Status          11239 non-null  int64
 7   State           11239 non-null  object
 8   Zone            11239 non-null  object
 9   Occupation      11239 non-null  object
 10  Product_Category  11239 non-null  object
 11  Orders          11239 non-null  int64
 12  Amount          11239 non-null  int32
dtypes: int32(1), int64(4), object(8)
memory usage: 1.2+ MB
```

In [27]:
```
data.head()
```

Out[27]:

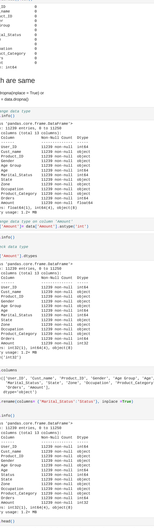
	User_ID	Cust_name	Product_ID	Gender	Age Group	Age	Status	State	Zone	Occupation	Product_Category	Orders	Amount
0	1002903	Sanskriti	P00125942	F	26-35	28	0	Maharashtra	Western	Healthcare	Auto	1	23952
1	1000732	Kartik	P00110942	F	26-35	35	1	Andhra Pradesh	Southern	Govt	Auto	3	23934
2	1001990	Bindu	P00110942	F	26-35	35	1	Uttar Pradesh	Central	Automobile	Auto	3	23924
3	1001425	Sudevi	P00237842	M	0-17	16	0	Karnataka	Southern	Construction	Auto	2	23912
4	1000588	Joni	P00057942	M	26-35	28	1	Gujarat	Western	Food Processing	Auto	2	23877

Age

In [28]:
```
ax= sns.countplot(data = data,x ='Age Group',hue ='Gender') # hue means side corner  value compair

for bars in ax.containers:
    ax.bar_label(bars)
```

Out[28]:

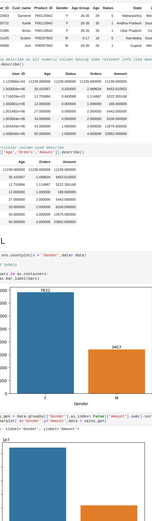


In [29]:
```
sales_age =data.groupby(['Age Group'],as_index= False)['Amount'].sum().sort_values(by='Amount',ascending = False)

sns.barplot(x = 'Age Group',y='Amount', data= sales_age)

#bars.bar_label('Age Group', ylabel='Amount')
```

Out[29]:



In [30]:
```
data.head()
```

Out[30]:

	User_ID	Cust_name	Product_ID	Gender	Age Group	Age	Status	State	Zone	Occupation	Product_Category	Orders	Amount
0	1002903	Sanskriti	P00125942	F	26-35	28	0	Maharashtra	Western	Healthcare	Auto	1	23952
1	1000732	Kartik	P00110942	F	26-35	35	1	Andhra Pradesh	Southern	Govt	Auto	3	23934
2	1001990	Bindu	P00110942	F	26-35	35	1	Uttar Pradesh	Central	Automobile	Auto	3	23924
3	1001425	Sudevi	P00237842	M	0-17	16	0	Karnataka	Southern	Construction	Auto	2	23912
4	1000588	Joni	P00057942	M	26-35	28	1	Gujarat	Western	Food Processing	Auto	2	23877

In [31]:
```
data.columns
```

Out[31]:
```
Index(['User_ID', 'Cust_name', 'Product_ID', 'Gender', 'Age Group', 'Age',
       'Status', 'State', 'Zone', 'Occupation', 'Product_Category', 'Orders',
       'Amount'],
      dtype='object')
```

State

In [32]:
```
sales_state = data.groupby(['State'],as_index= False)['Orders'].sum().sort_values(by='Orders', ascending = False).head(10)

sns.set(rc={'figure.figsize':(20,5)})
sns.barplot(data=sales_state,x = 'State',y='Orders')

#bars.bar_label('State', ylabel='Orders')
```

Out[32]:

In [ ]:
```

```