```
# importing libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
# importing dataset
data = pd.read csv("Diwali Sales Data.csv", encoding='unicode escape')
# top 5 and bottom 5 rows of dataset
data.head(5)
   User ID Cust name Product ID Gender Age Group Age
                                                       Marital Status
  1002903 Sanskriti P00125942
                                           26-35
                                                                    0
0
                                                   28
1 1000732
               Kartik P00110942
                                           26-35
                                                   35
                                                                    1
2 1001990
               Bindu P00118542
                                           26-35
                                                   35
                                                                    1
               Sudevi P00237842
3 1001425
                                            0-17
                                                   16
4 1000588
                Joni P00057942
                                           26-35
                                                   28
                                                                     1
                                     М
                                 Occupation Product Category Orders
           State
                      Zone
0
     Maharashtra
                                  Healthcare
                   Western
                                                                   1
                                                        Auto
1 Andhra Pradesh
                  Southern
                                       Govt
                                                        Auto
                                                                   3
2
   Uttar Pradesh
                  Central
                                  Automobile
                                                        Auto
                                                                   3
3
       Karnataka Southern
                               Construction
                                                        Auto
                                                                   2
                                                                   2
          Gujarat
                   Western Food Processing
                                                        Auto
   Amount
           Status
                   unnamed1
  23952.0
               NaN
0
                        NaN
1
  23934.0
               NaN
                        NaN
   23924.0
               NaN
                        NaN
3
  23912.0
               NaN
                        NaN
  23877.0
              NaN
                        NaN
data.tail(5)
       User ID
                 Cust_name Product_ID Gender Age Group Age
Marital Status
11246 1000695
                   Manning P00296942
                                           М
                                                 18-25
                                                         19
1
```

```
11247 1004089 Reichenbach P00171342
                                         М
                                              26-35
                                                      33
0
11248
     1001209
                    Oshin P00201342
                                         F
                                              36-45
                                                      40
11249
     1004023
                   Noonan P00059442
                                               36-45
                                                      37
11250
     1002744
                  Brumley P00281742
                                               18-25
                                                      19
                                Occupation Product_Category Orders
               State
                         Zone
Amount \
11246
         Maharashtra
                      Western
                                 Chemical
                                                   Office
                                                                4
370.0
11247
             Haryana Northern
                               Healthcare
                                                Veterinary
                                                                3
367.0
11248
      Madhya Pradesh Central
                                  Textile
                                                   Office
                                                                4
213.0
11249
           Karnataka Southern Agriculture
                                                   Office
                                                                3
206.0
11250
         Maharashtra Western
                               Healthcare
                                                   Office
                                                                3
188.0
      Status
              unnamed1
11246
         NaN
                  NaN
11247
         NaN
                  NaN
11248
         NaN
                  NaN
         NaN
                  NaN
11249
11250
         NaN
                  NaN
# no of rows and columns
data.shape
print(f'Number of rows in dataset:{data.shape[0]}')
print(f'Number of columns in dataset:{data.shape[1]}')
Number of rows in dataset:11251
Number of columns in dataset:15
# Display columns
data.columns
Index(['User_ID', 'Cust_name', 'Product_ID', 'Gender', 'Age Group',
'Age',
      'Marital_Status', 'State', 'Zone', 'Occupation',
dtype='object')
# stats about dataset
data.describe(include='all')
```

	l	lser_ID (Cust_name	Product	t_ID	Gende	r Ag	je Group		
Age \	1 1051	0004	11051	1.1	1251	1100	1	11051		
count		.00e+04	11251	1.	1251	1125	L	11251		
11251.00	00000	NaN	1250	-)) E 1		,	7		
unique		NaN	1250	4	2351	4	2	7		
NaN		N - N	مطاعات أحالا	DOOGE	- 2 4 2	ı	-	26 25		
top		NaN	Vishakha	P00265	0242	l	=	26-35		
NaN		NI NI	42		F 2	7043	•	45.40		
freq		NaN	42		53	7842	<u> </u>	4543		
NaN	1 0000	00400	NI - NI		NI - NI	N1 - N		NI - NI		
mean		04e+06	NaN		NaN	Nal	N	NaN		
35.4212		25 02	N. N.							
std		.25e+03	NaN		NaN	Nal	N	NaN		
12.7541										
min		01e+06	NaN		NaN	Nal	N	NaN		
12.0000		100 00								
25%		192e+06	NaN		NaN	Nal	N	NaN		
27.0000										
50%)65e+06	NaN		NaN	Nal	N	NaN		
33.0000										
75%		130e+06	NaN		NaN	Nal	N	NaN		
43.0000										
max	1.0060)40e+06	NaN		NaN	Nal	١	NaN		
max	1.0060)40e+06	NaN		NaN	Nal	١	NaN		
max	1.0060 00			State	NaN					
max 92.0000	1.0060 00 Marita	nl_Statu		State	NaN			NaN upation		
max 92.0000 Product	1.0060 00 Marita _Catego	al_Statu ory \	5			Zone (ıpation		
max 92.0000 Product count	1.0060 00 Marita _Catego	nl_Statu	5	State 11251						
max 92.0000 Product count 11251	1.0060 00 Marita _Catego	al_Statu ory \ 51.00000	5	11251		Zone (11251		upation 11251		
max 92.0000 Product count 11251 unique	1.0060 00 Marita _Catego	al_Statu ory \	5			Zone (ıpation		
max 92.0000 Product count 11251 unique 18	1.0060 00 Marita _Catego	al_Statu ory \ 51.00000 Na	5 9 N	11251 16	1	Zone (11251 5)ccu	11251 15	Clothing	
max 92.00000 Product count 11251 unique 18 top	1.0060 00 Marita _Catego 1125	al_Statu ory \ 51.00000	5 9 N	11251	1	Zone (11251)ccu	upation 11251	Clothing	J
max 92.00000 Product count 11251 unique 18 top Apparel	1.0060 00 Marita _Catego 1125	al_Statu ory \ 51.00000 Nal	s 9 N Uttar	11251 16 Pradesh	1	Zone (11251 5 ntral)ccu	11251 15 Sector	Clothing)
max 92.00000 Product count 11251 unique 18 top Apparel freq	1.0060 00 Marita _Catego 1125	al_Statu ory \ 51.00000 Na	s 9 N Uttar	11251 16	1	Zone (11251 5)ccu	11251 15	Clothing	J
max 92.00000 Product count 11251 unique 18 top Apparel freq 2655	1.0060 00 Marita _Catego 1125	al_Statu ory \ 51.00000 Nal Nal	s 9 N N Uttar	11251 16 Pradesh 1946	1	Zone (11251 5 ntral 4296)ccu	11251 15 Sector 1588	Clothing	j
max 92.00000 Product count 11251 unique 18 top Apparel freq 2655 mean	1.0060 00 Marita _Catego 1125	al_Statu ory \ 51.00000 Nal	s 9 N N Uttar	11251 16 Pradesh	1	Zone (11251 5 ntral)ccu	11251 15 Sector	Clothing)
max 92.00000 Product count 11251 unique 18 top Apparel freq 2655 mean NaN	1.0060 00 Marita _Catego 1125	al_Statu ory \ 51.00000 Nal Nal 0.42031	s O N Uttar N	11251 16 Pradesh 1946 NaN	1	Zone (11251 5 ntral 4296 NaN)ccu	11251 15 Sector 1588 NaN	Clothing	3
max 92.0000 Product count 11251 unique 18 top Apparel freq 2655 mean NaN std	1.0060 00 Marita _Catego 1125	al_Statu ory \ 51.00000 Nal Nal	s O N Uttar N	11251 16 Pradesh 1946	1	Zone (11251 5 ntral 4296)ccu	11251 15 Sector 1588	Clothing	J
max 92.00000 Product count 11251 unique 18 top Apparel freq 2655 mean NaN std NaN	1.0060 00 Marita _Catego 1125	nl_Status ory \ 51.000000 Nal Nal 0.420318	s 9 N Uttar N 8	11251 16 Pradesh 1946 NaN	1	Zone (11251 5 ntral 4296 NaN NaN)ccu	11251 15 Sector 1588 NaN NaN	Clothing	J
max 92.00000 Product count 11251 unique 18 top Apparel freq 2655 mean NaN std NaN min	1.0060 00 Marita _Catego 1125	al_Statu ory \ 51.00000 Nal Nal 0.42031	s 9 N Uttar N 8	11251 16 Pradesh 1946 NaN	1	Zone (11251 5 ntral 4296 NaN)ccu	11251 15 Sector 1588 NaN	Clothing)
max 92.00000 Product count 11251 unique 18 top Apparel freq 2655 mean NaN std NaN min NaN	1.0060 00 Marita _Catego 1125	Nal 0.42031 0.49363	s O N Uttar N B 2	11251 16 Pradesh 1946 NaN NaN	1	Zone (11251 5 ntral 4296 NaN NaN)ccu	pation 11251 15 Sector 1588 NaN NaN NaN	Clothing	3
max 92.00000 Product count 11251 unique 18 top Apparel freq 2655 mean NaN std NaN min NaN 25%	1.0060 00 Marita _Catego 1125	nl_Status ory \ 51.000000 Nal Nal 0.420318	s O N Uttar N B 2	11251 16 Pradesh 1946 NaN	1	Zone (11251 5 ntral 4296 NaN NaN)ccu	11251 15 Sector 1588 NaN NaN	Clothing)
max 92.00000 Product count 11251 unique 18 top Apparel freq 2655 mean NaN std NaN min NaN 25% NaN	1.0060 00 Marita _Catego 1125	Nal 0.420313 0.493633 0.00000	5 9 N N Uttar N 3 2 9	11251 16 Pradesh 1946 NaN NaN NaN	1	Zone (11251 5 ntral 4296 NaN NaN NaN)ccu	pation 11251 15 Sector 1588 NaN NaN NaN NaN	Clothing	J
max 92.00000 Product count 11251 unique 18 top Apparel freq 2655 mean NaN std NaN min NaN 25% NaN 50%	1.0060 00 Marita _Catego 1125	Nal 0.42031 0.49363	5 9 N N Uttar N 3 2 9	11251 16 Pradesh 1946 NaN NaN	1	Zone (11251 5 ntral 4296 NaN NaN)ccu	pation 11251 15 Sector 1588 NaN NaN NaN	Clothing)
max 92.00000 Product count 11251 unique 18 top Apparel freq 2655 mean NaN std NaN min NaN 25% NaN 50% NaN	1.0060 00 Marita _Catego 1125	Nal Nal 0.420313 0.493633 0.000000	5 0 N Uttar N 3 2 0 0	11251 16 Pradesh 1946 NaN NaN NaN	1	Zone (11251 5 ntral 4296 NaN NaN NaN NaN)ccu	spation 11251 15 Sector 1588 NaN NaN NaN NaN NaN	Clothing)
max 92.00000 Product count 11251 unique 18 top Apparel freq 2655 mean NaN std NaN min NaN 25% NaN 50% NaN 75%	1.0060 00 Marita _Catego 1125	Nal 0.420313 0.493633 0.00000	5 0 N Uttar N 3 2 0 0	11251 16 Pradesh 1946 NaN NaN NaN	1	Zone (11251 5 ntral 4296 NaN NaN NaN)ccu	pation 11251 15 Sector 1588 NaN NaN NaN NaN	Clothing)
max 92.00000 Product count 11251 unique 18 top Apparel freq 2655 mean NaN std NaN min NaN 25% NaN 50% NaN 75% NaN	1.0060 00 Marita _Catego 1125	Al_Status Ory \ 61.000000 Nal Nal 0.420313 0.493633 0.000000 0.000000 1.000000	5 9 N N Uttar N 3 2 9	11251 16 Pradesh 1946 NaN NaN NaN NaN	1	Zone (11251 5 ntral 4296 NaN NaN NaN NaN)ccu	pation 11251 15 Sector 1588 NaN NaN NaN NaN NaN	Clothing	Ð
max 92.00000 Product count 11251 unique 18 top Apparel freq 2655 mean NaN	1.0060 00 Marita _Catego 1125	Nal Nal 0.420313 0.493633 0.000000	5 9 N N Uttar N 3 2 9	11251 16 Pradesh 1946 NaN NaN NaN	1	Zone (11251 5 ntral 4296 NaN NaN NaN NaN)ccu	spation 11251 15 Sector 1588 NaN NaN NaN NaN NaN	Clothing)

```
0rders
                                      Status
                             Amount
                                              unnamed1
        11251.000000
                       11239.000000
count
                                         0.0
                                                    0.0
                                                    NaN
unique
                                         NaN
                  NaN
                                 NaN
                  NaN
                                NaN
                                         NaN
                                                    NaN
top
                  NaN
                                NaN
                                         NaN
                                                    NaN
freq
            2.489290
                        9453.610858
                                         NaN
                                                    NaN
mean
std
            1.115047
                        5222.355869
                                         NaN
                                                    NaN
            1.000000
min
                         188.000000
                                         NaN
                                                    NaN
25%
            1.500000
                        5443.000000
                                         NaN
                                                    NaN
50%
            2.000000
                        8109.000000
                                         NaN
                                                    NaN
                       12675.000000
75%
            3.000000
                                         NaN
                                                    NaN
max
            4.000000
                     23952.000000
                                         NaN
                                                    NaN
# infromation about dataset
data.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11251 entries, 0 to 11250
Data columns (total 15 columns):
                        Non-Null Count
     Column
                                         Dtype
     -----
     User ID
 0
                                         int64
                        11251 non-null
 1
     Cust name
                        11251 non-null
                                         object
 2
                        11251 non-null
     Product ID
                                         object
 3
     Gender
                        11251 non-null
                                         object
 4
     Age Group
                        11251 non-null
                                         object
 5
                        11251 non-null
                                         int64
     Age
 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
     0rders
                        11251 non-null
                                         int64
                        11239 non-null
 12
    Amount
                                         float64
 13
     Status
                        0 non-null
                                         float64
                        0 non-null
14
     unnamed1
                                         float64
dtypes: float64(3), int64(4), object(8)
memory usage: 1.3+ MB
# checking if any column is null
data.isnull().sum()
User ID
                         0
                         0
Cust name
                         0
Product ID
Gender
                         0
                         0
Age Group
Age
                         0
                         0
Marital Status
```

```
State
                       0
                       0
Zone
Occupation
                       0
Product Category
                       0
0rders
                       0
Amount
                       12
Status
                    11251
                    11251
unnamed1
dtype: int64
# handling null vaues
average = data.Amount.mean()
average
np.float64(9453.610857727557)
data.Amount.fillna(round(average, 0), inplace=True)
# remove Status and unnamed1 columns
data.drop(columns=['Status', 'unnamed1'],inplace=True)
data.columns
Index(['User_ID', 'Cust_name', 'Product_ID', 'Gender', 'Age Group',
       'Marital Status', 'State', 'Zone', 'Occupation',
dtype='object')
# check whether duplicate values or not
data.duplicated().sum()
np.int64(8)
# handling dupicate values
data[data.duplicated()==True]
       User ID
                Cust name Product ID Gender Age Group Age
Marital Status
19
       1001883
                   Praneet P00029842
                                                 51-55
                                                        54
                                          М
1
4404
      1004725
                   Jackson P00150842
                                                 36-45
                                                        37
5703
       1003208
                    Bowman P00171642
                                                 26-35
                                                        31
5908
                                                 26-35
       1001260
                   Dheerai P00344042
                                          М
                                                        28
6173
       1001325
                     Reese
                           P00111742
                                                 26-35
                                                        27
1
       1000083
                     Gute P00242842
                                                 26-35
8651
                                          М
                                                        35
```

0	1001476	A	.d	DOOO	26042	M	10.25	25	
8941 0	1001476	Anu	ideep	P000	36842	М	18-25	25	
10571 1	1004404	Ritten	ibach	P001	50142	F	26-35	28	
0 1	,	State	Zo	ne	0ccupa	ation	Produ	ct_(Category
Orders 19 1	\ Uttar Pr	adesh	Cent	al	Hospita	ality			Auto
4404 4	Mahara	shtra	Weste	ern	Hospita	ality	Electronic	s &	Gadgets
5703 4		Bihar	East	ern	Agricul	lture	Electronic	s &	Gadgets
5908 4	Mahara	shtra	Weste	ern	IT Se	ector	Electronic	s &	Gadgets
6173 3	Gu	ıjarat	Weste	ern	Construc	ction	Electronic	s &	Gadgets
8651 3	Uttar Pr	adesh	Cent	al	Hospita	ality	Clothin	g &	Apparel
8941	Mahara	shtra	Weste	ern	IT Se	ector	Clothin	g &	Apparel
4 10571 3	На	ryana	Northe	ern	Avia	ation	Electronic	s &	Gadgets
19 4404 5703 5908 6173 8651 8941 10571	Amount 23568.0 9859.0 8088.0 8015.0 7923.0 5345.0 5202.0 2304.0								
data.d	rop_dupli	.cates()							
Marita	User_ID l Status	Cust	_name	Prod	uct_ID (Gender	Age Group	Age	e
0	1002903	-	kriti	P00	125942	F	26-35	28	3
1	1000732	K	Cartik	P00	110942	F	26-35	35	5
2	1001990		Bindu	P00	118542	F	26-35	35	5
1	1001425	S	udevi	P00	237842	М	0-17	16	5
0 4 1	1000588		Joni	P00	057942	М	26-35	28	3

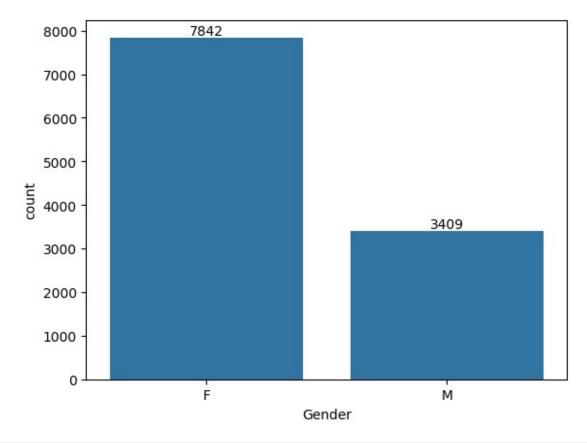
11246	1000695	Ma	nning	P00	296942	М	18-25	19
1 11247	1004089	Reiche	nbach	P00	171342	М	26-35	33
0 11248	1001209		0shin	P00	201342	F	36-45	40
0 11249	1004023	s N	oonan	P00	059442	М	36-45	37
0 11250	1002744		umley		281742	F	18-25	19
0	1002744	. Бі	unicey	100	201742	ı	10-25	19
		State	Z	one	0cc	upation	Product_	_Category
Orders 0	\ Maha	rashtra	West	ern	Hea	lthcare		Auto
1 1	Andhra	Pradesh	South	ern		Govt		Auto
3 2	llttar	Pradesh	Cent		Διι+	omobile		Auto
3								
3 3 2		rnataka	South			ruction		Auto
4 2		Gujarat	West	ern	Food Pro	cessing		Auto
11246 4	Maha	rashtra	West	ern	C	hemical		Office
11247		Haryana	North	ern	Hea	lthcare	Ve	eterinary
3 11248 4	Madhya	Pradesh	Cent	ral		Textile		Office
11249 3	Ka	rnataka	South	ern	Agri	culture		Office
11250 3	Maha	rashtra	West	ern	Неа	lthcare		Office
0 1 2 3 4 11246 11247 11248 11249 11250	Amount 23952.0 23934.0 23924.0 23912.0 23877.0 370.0 367.0 213.0 206.0 188.0							

```
[11243 rows x 13 columns]
# datatypes of columns
data.dtypes
User ID
                       int64
Cust name
                      object
Product ID
                      object
Gender
                      object
Age Group
                      object
                       int64
Age
Marital_Status
                       int64
State
                      object
Zone
                      object
Occupation
                      object
Product Category
                      object
0rders
                       int64
Amount
                     float64
dtype: object
# Amount datatype convert from float64 to int64
data.Amount=data.Amount.astype('int64')
data.dtypes
User ID
                      int64
                     object
Cust name
Product ID
                     object
Gender
                     object
Age Group
                     object
                     int64
Age
                     int64
Marital_Status
State
                     object
Zone
                     object
Occupation
                     object
Product Category
                     object
0rders
                      int64
Amount
                      int64
dtype: object
# rename columns
data.rename(columns={'Marital Status':'Married Or Not'},inplace=True)
data.columns
Index(['User ID', 'Cust name', 'Product ID', 'Gender', 'Age Group',
'Age',
       'Married Or Not', 'State', 'Zone', 'Occupation',
'Product Category',
       'Orders', 'Amount'],
      dtype='object')
```

Exploratory Data Analysis

Gender

```
count_plot = sns.countplot(x='Gender',data=data,stat='count')
for bars in count_plot.containers:
    count_plot.bar_label(bars)
```

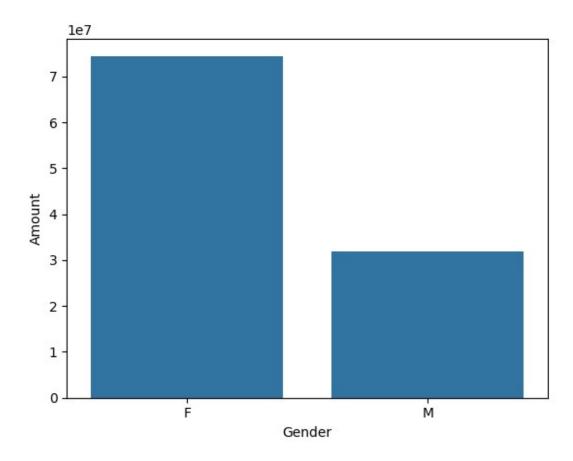


```
# genderwise amount
am =data.groupby(['Gender'],as_index=False)
['Amount'].sum().sort_values(by='Amount',ascending=False)
am

Gender Amount
0 F 74430393
1 M 31932184

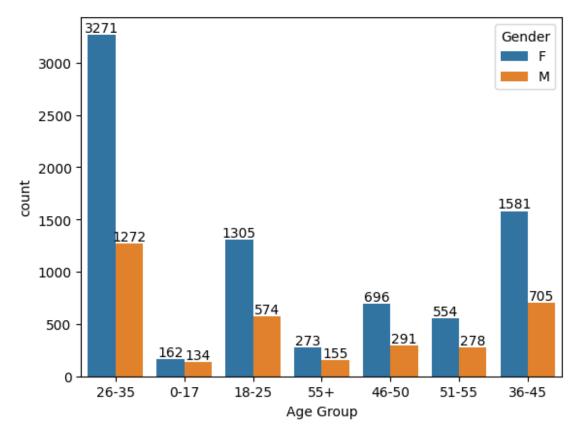
sns.barplot(data=am,x= 'Gender',y='Amount')

<Axes: xlabel='Gender', ylabel='Amount'>
```



Age_Group

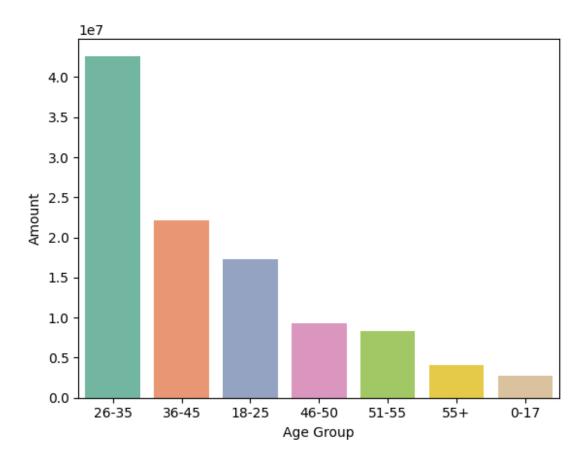
```
age = sns.countplot(data = data, x = 'Age Group', hue = 'Gender')
for bars in age.containers:
    age.bar_label(bars)
```



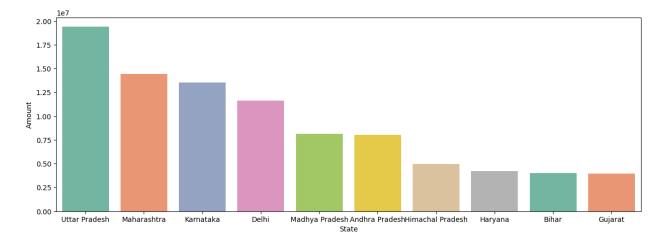
```
import warnings
warnings.filterwarnings('ignore')

# age_group wise amount
am1 = data.groupby(['Age Group'],as_index=False)
['Amount'].sum().sort_values(by='Amount',ascending=False)
sns.barplot(data=am1,x='Age Group',y= 'Amount', palette='Set2')

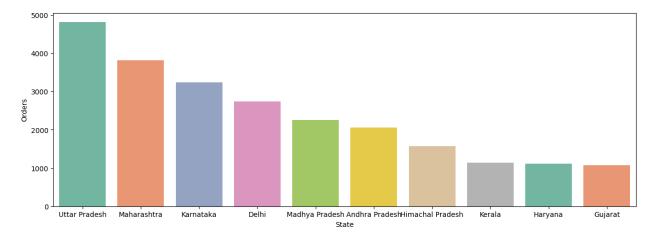
<Axes: xlabel='Age Group', ylabel='Amount'>
```



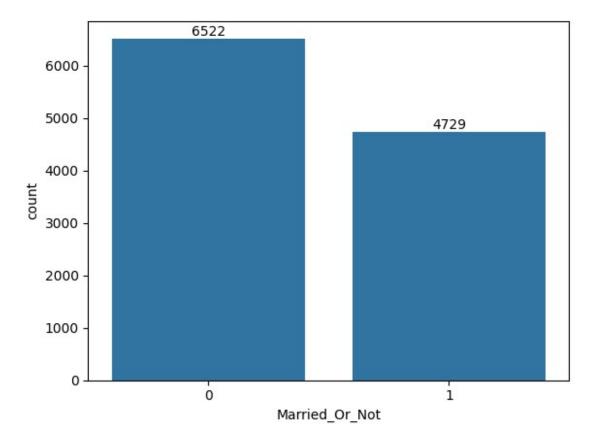
```
# top 10 states-wise total amount
states =data.groupby(['State'],as_index=False)
['Amount'].sum().sort_values(by= 'Amount',ascending=False).head(10)
plt.figure(figsize=(15,5))
sns.barplot(data= states,x='State',y='Amount',palette='Set2')
plt.show()
```



```
# top 10 state wise orders
# top 10 states
states =data.groupby(['State'],as_index=False)
['Orders'].sum().sort_values(by= 'Orders',ascending=False).head(10)
plt.figure(figsize=(15,5))
sns.barplot(data= states,x='State',y='Orders',palette='Set2')
plt.show()
```



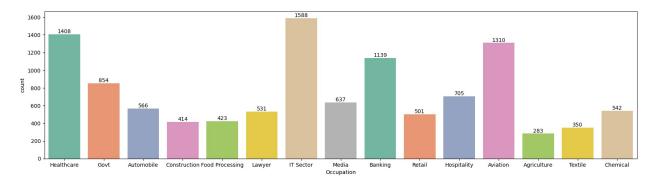
```
married =sns.countplot(data = data, x = 'Married_Or_Not')
plt.figure(figsize=(4,5))
for bars in married.containers:
    married.bar_label(bars)
```



```
<Figure size 400x500 with 0 Axes>
data.Married_Or_Not.unique()
array([0, 1])
```

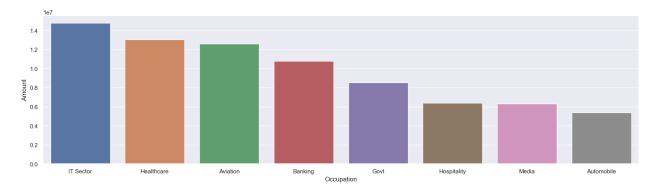
Occupation

```
plt.figure(figsize=(20,5))
ax = sns.countplot(data = data, x = 'Occupation',palette='Set2')
for bars in ax.containers:
    ax.bar_label(bars)
```



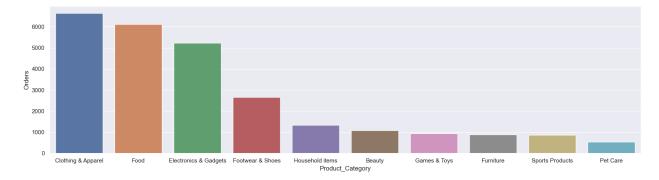
```
# top 8 performing occupation-wise total amount
top_states = data.groupby(['Occupation'], as_index=False)
['Amount'].sum().sort_values(by='Amount', ascending=False).head(8)
sns.set(rc={'figure.figsize':(20,5)})
sns.barplot(data = top_states, x = 'Occupation',y=
'Amount',palette='deep')

<a href="Axes: xlabel='Occupation'">Axes: xlabel='Occupation'</a>, ylabel='Amount'>
```



```
# top 10 most sold products
top_products = data.groupby(['Product_Category'], as_index=False)
['Orders'].sum().sort_values(by='Orders', ascending=False).head(10)
sns.set(rc={'figure.figsize':(20,5)})
sns.barplot(data = top_products, x = 'Product_Category',y='Orders',palette='deep')

<Axes: xlabel='Product_Category', ylabel='Orders'>
```



Conclusion

During Dwali Sales, 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