In [ ]:

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

In [ ]:

```
df=pd.read_csv(r"C:\Users\DELL\Downloads\archive (2) (1)\Online Shop Customer Sales
df
```

Out[ ]:

	Customer_id	Age	Gender	Revenue_Total	N_Purchases	Purchase_DATE	Purchase_
0	504308	53	0	45.3	2	22.06.21	
1	504309	18	1	36.2	3	10.12.21	
2	504310	52	1	10.6	1	14.03.21	
3	504311	29	0	54.1	5	25.10.21	
4	504312	21	1	56.9	1	14.09.21	
...	...	...	...	...	...	...	
65791	570099	30	1	10.9	4	15.02.21	
65792	570100	33	0	29.3	1	05.02.21	
65793	570101	50	0	25.4	5	29.10.21	
65794	570102	56	0	29.2	1	09.12.21	
65795	570103	25	0	5.3	2	23.10.21	

65796 rows × 12 columns

In [ ]:

```
df.loc[df["Gender"]==1,"Gender"]="Female"
df.loc[df["Gender"]==0, "Gender"]="Male"
```

In [ ]:

```
df.loc[df["Pay_Method"]==0, "Pay_Method"]="Digital Wallets"
df.loc[df["Pay_Method"]==1,"Pay_Method"]="Card"
df.loc[df["Pay_Method"]==2, "Pay_Method"]="PayPal"
df.loc[df["Pay_Method"]==3, "Pay_Method"]="Other"
```

In [ ]:

```
df.loc[df["Browser"]==0,"Browser"]="Chrome"
df.loc[df["Browser"]==1,"Browser"]="Safari"
df.loc[df["Browser"]==2,"Browser"]="Edge"
df.loc[df["Browser"]==3,"Browser"]="Other"
```

In [ ]:

```
df.loc[df["Newsletter"]==0,"Newsletter"]="not subscribed"
df.loc[df["Newsletter"]==1,"Newsletter"]="subscribed"
```

In [ ]:

```
df.loc[df["Voucher"]==0,"Voucher"]="Not_Used"
df.loc[df["Voucher"]==1,"Voucher"]="Used"
```

```
In [ ]: df
```

Out[ ]:

	Customer_id	Age	Gender	Revenue_Total	N_Purchases	Purchase_DATE	Purchase_
0	504308	53	Male	45.3	2	22.06.21	
1	504309	18	Female	36.2	3	10.12.21	
2	504310	52	Female	10.6	1	14.03.21	
3	504311	29	Male	54.1	5	25.10.21	
4	504312	21	Female	56.9	1	14.09.21	
...	...	...	...	...	...	...	
65791	570099	30	Female	10.9	4	15.02.21	
65792	570100	33	Male	29.3	1	05.02.21	
65793	570101	50	Male	25.4	5	29.10.21	
65794	570102	56	Male	29.2	1	09.12.21	
65795	570103	25	Male	5.3	2	23.10.21	

65796 rows × 12 columns

```
In [ ]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 65796 entries, 0 to 65795
Data columns (total 12 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   Customer_id     65796 non-null  int64  
 1   Age             65796 non-null  int64  
 2   Gender          65796 non-null  object 
 3   Revenue_Total   65796 non-null  float64
 4   N_Purchases     65796 non-null  int64  
 5   Purchase_DATE   65796 non-null  object 
 6   Purchase_VALUE  65796 non-null  float64
 7   Pay_Method      65796 non-null  object 
 8   Time_Spent      65796 non-null  int64  
 9   Browser         65796 non-null  object 
 10  Newsletter      65796 non-null  object 
 11  Voucher         65796 non-null  object 
dtypes: float64(2), int64(4), object(6)
memory usage: 6.0+ MB
```

In [ ]:

```python
def age_group(age):
    if(age<20):
        return "20-25"
    elif(age<25):
        return "25-30"
    elif(age<30):
        return "30-35"
    elif(age<35):
        return "35-40"
    elif(age<40):
        return "40-45"
    elif(age<45):
        return "45-50"
    else:
        return "Above 50"
```

In [ ]:

```python
df["Age_Group"]=df["Age"].apply(age_group)
df
```

Out[ ]:

	Customer_id	Age	Gender	Revenue_Total	N_Purchases	Purchase_DATE	Purchase_
0	504308	53	Male	45.3	2	22.06.21	
1	504309	18	Female	36.2	3	10.12.21	
2	504310	52	Female	10.6	1	14.03.21	
3	504311	29	Male	54.1	5	25.10.21	
4	504312	21	Female	56.9	1	14.09.21	
...	...	...	...	...	...	...	
65791	570099	30	Female	10.9	4	15.02.21	
65792	570100	33	Male	29.3	1	05.02.21	
65793	570101	50	Male	25.4	5	29.10.21	
65794	570102	56	Male	29.2	1	09.12.21	
65795	570103	25	Male	5.3	2	23.10.21	

65796 rows × 13 columns

```
In [ ]: df["Purchase_DATE"]=pd.to_datetime(df["Purchase_DATE"])
```

```
In [ ]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 65796 entries, 0 to 65795
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype         
---  ------          --------------  -----         
 0   Customer_id     65796 non-null  int64         
 1   Age             65796 non-null  int64         
 2   Gender          65796 non-null  object        
 3   Revenue_Total   65796 non-null  float64       
 4   N_Purchases     65796 non-null  int64         
 5   Purchase_DATE   65796 non-null  datetime64[ns]
 6   Purchase_VALUE  65796 non-null  float64       
 7   Pay_Method      65796 non-null  object        
 8   Time_Spent      65796 non-null  int64         
 9   Browser         65796 non-null  object        
 10  Newsletter      65796 non-null  object        
 11  Voucher         65796 non-null  object        
 12  Age_Group       65796 non-null  object        
dtypes: datetime64[ns](1), float64(2), int64(4), object(6)
memory usage: 6.5+ MB
```

In [ ]:

```
df.shape
```

Out[ ]:

```
(65796, 13)
```

In [ ]:

```
df.columns
```

Out[ ]:

```
Index(['Customer_id', 'Age', 'Gender', 'Revenue_Total', 'N_Purchases',
       'Purchase_DATE', 'Purchase_VALUE', 'Pay_Method', 'Time_Spent',
       'Browser', 'Newsletter', 'Voucher', 'Age_Group'],
      dtype='object')
```

In [ ]:

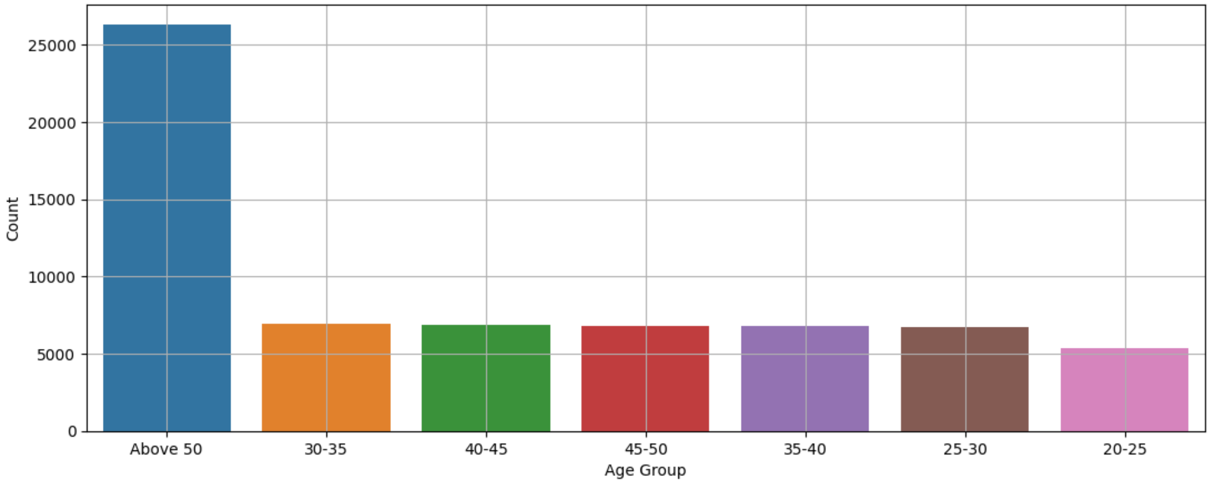
```
counts= df['Age_Group'].value_counts()
counts=counts.reset_index()
counts=counts.rename(columns={"index":"Age Group", "Age_Group":"Count"})
counts
```

Out[ ]:

	Age Group	Count
0	Above 50	26291
1	30-35	6906
2	40-45	6857
3	45-50	6827
4	35-40	6798
5	25-30	6741
6	20-25	5376

In [ ]:

```
plt.figure(figsize=(13,5))
sns.barplot(data=counts, x="Age Group", y="Count")
plt.grid()
```



What Age Group Buys from us the most ?

What Payment Method is used most by Age_Groups

```
In [ ]: ndf=df.groupby(["Pay_Method","Age_Group"]).agg(
            Use=("Pay_Method", "count")
        ).reset_index().sort_values(by="Use",ascending=False)
        ndf
```

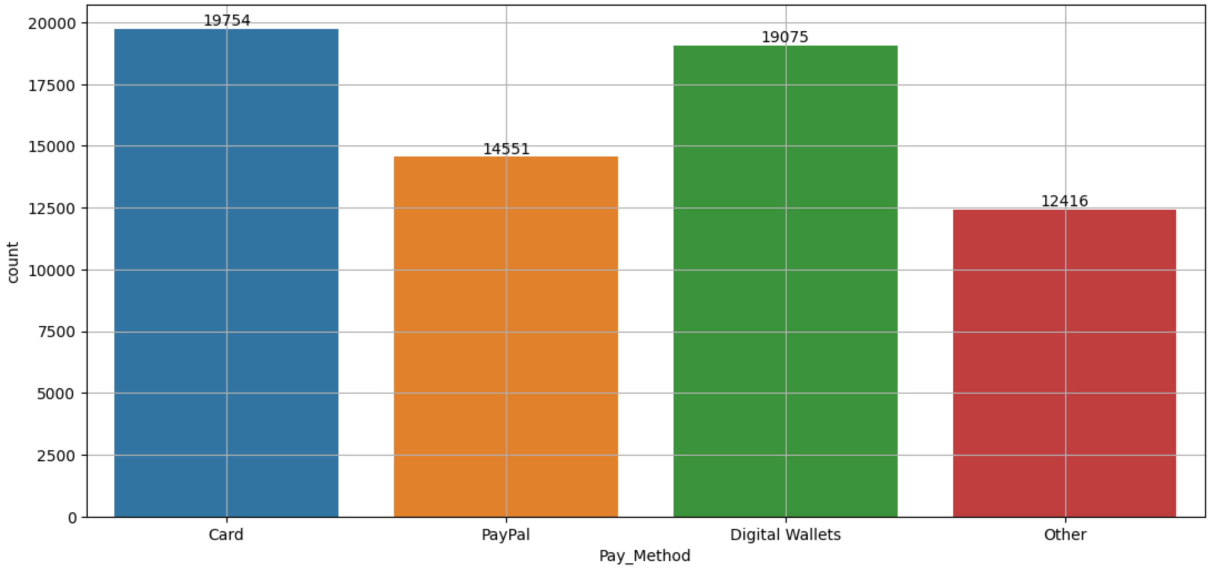
Out[ ]:

	Pay_Method	Age_Group	Use
6	Card	Above 50	7871
13	Digital Wallets	Above 50	7609
27	PayPal	Above 50	5876
20	Other	Above 50	4935
2	Card	30-35	2129
5	Card	45-50	2073
4	Card	40-45	2057
3	Card	35-40	2053
11	Digital Wallets	40-45	2018
1	Card	25-30	2005
8	Digital Wallets	25-30	1977
10	Digital Wallets	35-40	1976
9	Digital Wallets	30-35	1966
12	Digital Wallets	45-50	1945
7	Digital Wallets	20-25	1584
0	Card	20-25	1566
26	PayPal	45-50	1523
23	PayPal	30-35	1509
24	PayPal	35-40	1495
25	PayPal	40-45	1487
22	PayPal	25-30	1481
16	Other	30-35	1302
18	Other	40-45	1295
19	Other	45-50	1286
15	Other	25-30	1278
17	Other	35-40	1274
21	PayPal	20-25	1180
14	Other	20-25	1046

In [ ]:

```
plt.figure(figsize=(13,6))
a=sns.countplot(data=df, x="Pay_Method")
plt.grid()
```

```
for i in a.containers:
    a.bar_label(i)
```



What Browsers Our Customers use most ?

In [ ]:

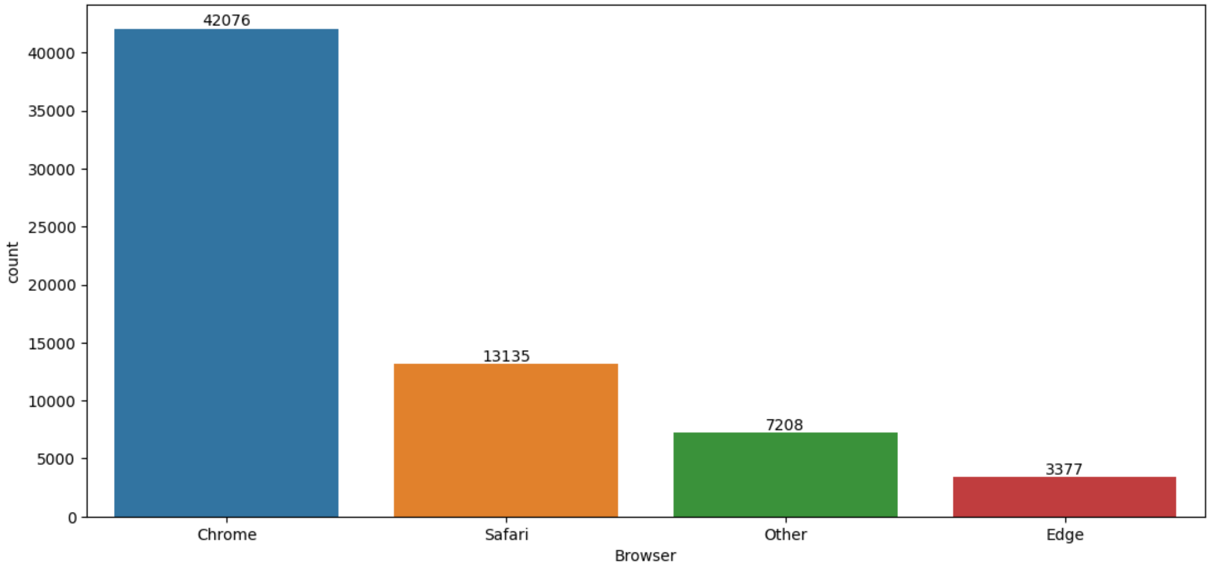
```
df.Browser.value_counts()
```

Out[ ]:

```
Chrome    42076
Safari    13135
Other      7208
Edge       3377
Name: Browser, dtype: int64
```

In [ ]:

```
plt.figure(figsize=(13,6))
b=sns.countplot(data=df, x="Browser")
for i in b.containers:
    b.bar_label(i)
```



In [ ]:

```
df["Month"]=df["Purchase_DATE"].dt.month_name()
```

```
df
```

Out[ ]:

	Customer_id	Age	Gender	Revenue_Total	N_Purchases	Purchase_DATE	Purchase_V
0	504308	53	Male	45.3	2	2021-06-22	
1	504309	18	Female	36.2	3	2021-10-12	
2	504310	52	Female	10.6	1	2021-03-14	
3	504311	29	Male	54.1	5	2021-10-25	
4	504312	21	Female	56.9	1	2021-09-14	
...	...	...	...	...	...	...	
65791	570099	30	Female	10.9	4	2021-02-15	
65792	570100	33	Male	29.3	1	2021-05-02	
65793	570101	50	Male	25.4	5	2021-10-29	
65794	570102	56	Male	29.2	1	2021-09-12	
65795	570103	25	Male	5.3	2	2021-10-23	

65796 rows × 14 columns

Highest Orders In Month

In [ ]:

```
df.Month.value_counts()
```

Out[ ]:

```
December     5643
January      5631
August       5625
May          5607
October      5563
July         5543
March        5467
June         5455
September    5447
April        5407
November     5330
February     5078
Name: Month, dtype: int64
```

Highest Sale In Month

In [ ]:

```
cdf=df.groupby("Month").agg(
    Sum=("Purchase_VALUE", "sum"),
    Avg=("Purchase_VALUE", "mean")
).reset_index()
cdf.sort_values(by="Sum",ascending=False)
```

Out[ ]:

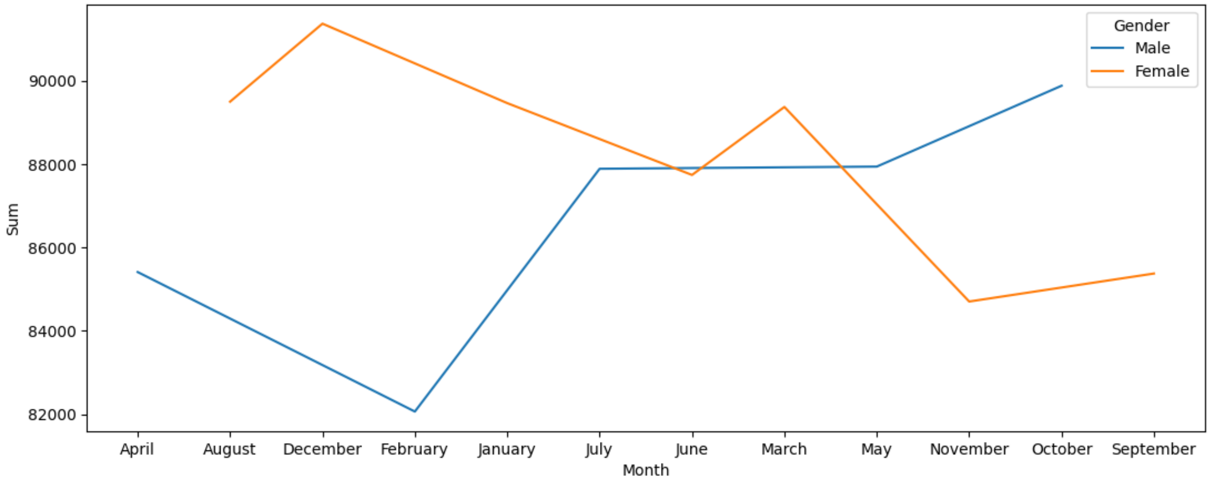
	Month	Sum	Avg
2	December	91364.573	16.190780
10	October	89876.060	16.156042
1	August	89494.439	15.910122
4	January	89459.438	15.886954
7	March	89367.929	16.346795
8	May	87938.008	15.683611
5	July	87886.125	15.855336
6	June	87737.820	16.083927
0	April	85409.355	15.796071
11	September	85374.270	15.673631
9	November	84703.236	15.891789
3	February	82065.269	16.160943

In [ ]:

```
plt.figure(figsize=(13,5))
sns.lineplot(data=cdf, x="Month", y="Sum",hue=df["Gender"])
```

Out[ ]:

```
<Axes: xlabel='Month', ylabel='Sum'>
```

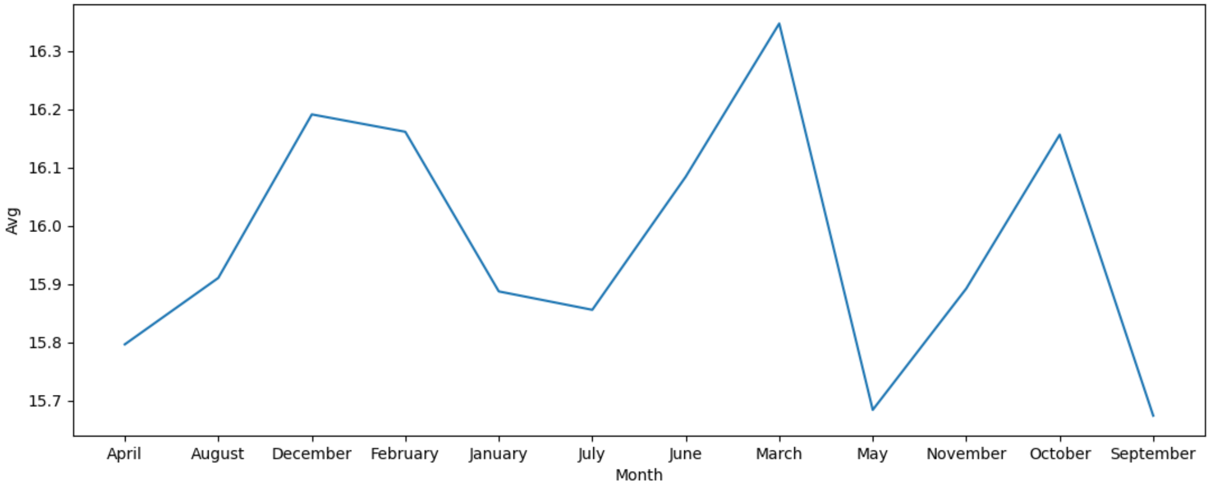


In [ ]:

```
plt.figure(figsize=(13,5))
sns.lineplot(data=cdf, x="Month", y="Avg")
```

Out[ ]:

```
<Axes: xlabel='Month', ylabel='Avg'>
```

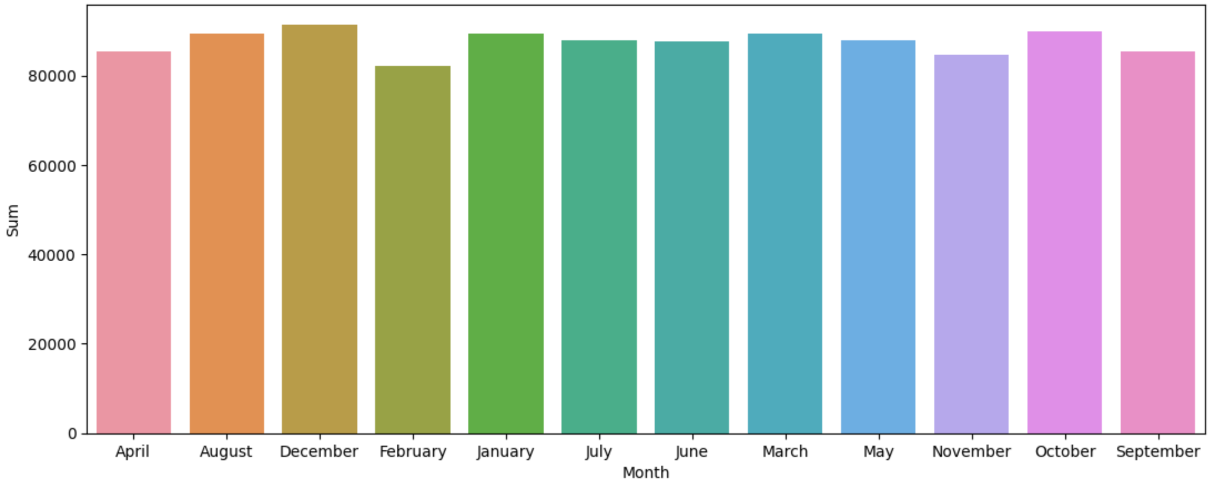


In [ ]:

```
plt.figure(figsize=(13,5))
sns.barplot(data=cdf, x="Month", y="Sum")
```

Out[ ]:

```
<Axes: xlabel='Month', ylabel='Sum'>
```



What is the average time spent by male and female respectively on our Website ?

In [ ]:

```
df.groupby("Gender")["Time_Spent"].mean()
```

Out[ ]:

```
Gender
Female    599.235647
Male      598.292268
Name: Time_Spent, dtype: float64
```

In [ ]:

```
# Not Much Difference!!!
```

Who are our Best Customers (Males Or Females ?)

In [ ]:

```
edf=df.groupby("Gender").agg(
        Total_Revenue = ('Revenue_Total','sum'),
        Average_Revenue = ('Revenue_Total','mean'),
        Number_of_Purchases = ('N_Purchases','mean'),
        Average_Time_Spent = ('Time_Spent','mean')
)
edf
```

Out[ ]:

	Total_Revenue	Average_Revenue	Number_of_Purchases	Average_Time_Spent
Gender				
Female	1224554.7	27.722419	3.994544	599.235647
Male	600161.5	27.754416	3.988254	598.292268