In [40]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt #visualizing data
%matplotlib inline
import seaborn as sns
```

In [41]:
```
df = pd.read_excel('Retail Sales Analysis_utf.xlsx')
df.shape
```

Out[41]: (2000, 11)

In [42]:
```
df.head()
```

Out[42]:

	transactions_id	sale_date	sale_time	customer_id	gender	age	category	quantiy	pric
0	180	2022-11-05	10:47:00	117	Male	41.0	Clothing	3.0	
1	522	2022-07-09	11:00:00	52	Male	46.0	Beauty	3.0	
2	559	2022-12-12	10:48:00	5	Female	40.0	Clothing	4.0	
3	1180	2022-01-06	08:53:00	85	Male	41.0	Clothing	3.0	
4	1522	2022-11-14	08:35:00	48	Male	46.0	Beauty	3.0	

In [43]:
```
df.info
```

Out[43]:
```
<bound method DataFrame.info of       transactions_id  sale_date sale_tim
e  customer_id  gender   age  \
0                180 2022-11-05  10:47:00          117    Male  41.0
1                522 2022-07-09  11:00:00           52    Male  46.0
2                559 2022-12-12  10:48:00            5  Female  40.0
3               1180 2022-01-06  08:53:00           85    Male  41.0
4               1522 2022-11-14  08:35:00           48    Male  46.0
...              ...        ...       ...          ...     ...   ...
1995            1857 2022-11-09  12:15:00          109    Male  60.0
1996             211 2022-09-12  14:02:00           54    Male  42.0
1997             650 2023-10-08  12:41:00           98    Male  55.0
1998            1211 2023-11-22  14:59:00           82    Male  42.0
1999            1650 2022-09-23  16:24:00           89    Male  55.0

         category  quantiy  price_per_unit   cogs  total_sale
0        Clothing      3.0           300.0  129.0       900.0
1          Beauty      3.0           500.0  145.0      1500.0
2        Clothing      4.0           300.0   84.0      1200.0
3        Clothing      3.0           300.0  129.0       900.0
4          Beauty      3.0           500.0  235.0      1500.0
...           ...      ...             ...    ...         ...
1995  Electronics      2.0            25.0    7.5        50.0
1996       Beauty      3.0           500.0  235.0      1500.0
1997  Electronics      1.0            30.0   15.0        30.0
1998       Beauty      3.0           500.0  235.0      1500.0
1999  Electronics      1.0            30.0   10.8        30.0

[2000 rows x 11 columns]>
```

In [44]:
```
pd.isnull(df).sum
```

Out[44]:
```
<bound method NDFrame._add_numeric_operations.<locals>.sum of       trans
actions_id  sale_date  sale_time  customer_id  gender    age  \
0              False      False      False        False   False  False
1              False      False      False        False   False  False
2              False      False      False        False   False  False
3              False      False      False        False   False  False
4              False      False      False        False   False  False
...              ...        ...        ...          ...     ...    ...
1995           False      False      False        False   False  False
1996           False      False      False        False   False  False
1997           False      False      False        False   False  False
1998           False      False      False        False   False  False
1999           False      False      False        False   False  False

      category  quantiy  price_per_unit   cogs  total_sale
0        False    False           False  False       False
1        False    False           False  False       False
2        False    False           False  False       False
3        False    False           False  False       False
4        False    False           False  False       False
...        ...      ...             ...    ...         ...
1995     False    False           False  False       False
1996     False    False           False  False       False
1997     False    False           False  False       False
1998     False    False           False  False       False
1999     False    False           False  False       False

[2000 rows x 11 columns]>
```

In [45]:
```
df.dropna(inplace=True)
```

In [46]:
```
df.shape
```

Out[46]:
```
(1987, 11)
```

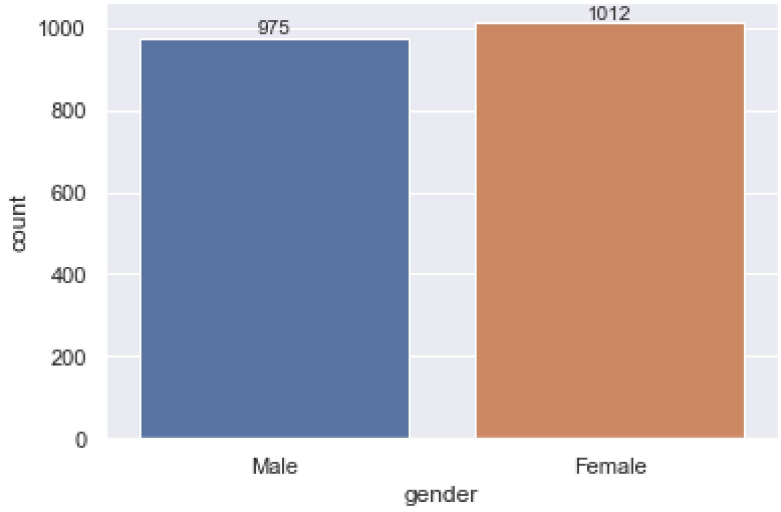
Exploratory Data Analysis

Gender

In [47]:
```
df.columns
```

Out[47]:
```
Index(['transactions_id', 'sale_date', 'sale_time', 'customer_id', 'gende
r',
       'age', 'category', 'quantiy', 'price_per_unit', 'cogs', 'total_sal
e'],
      dtype='object')
```

In [48]:
```
ax=sns.countplot(x='gender',data=df)
for bars in ax.containers:
    ax.bar_label(bars)
```

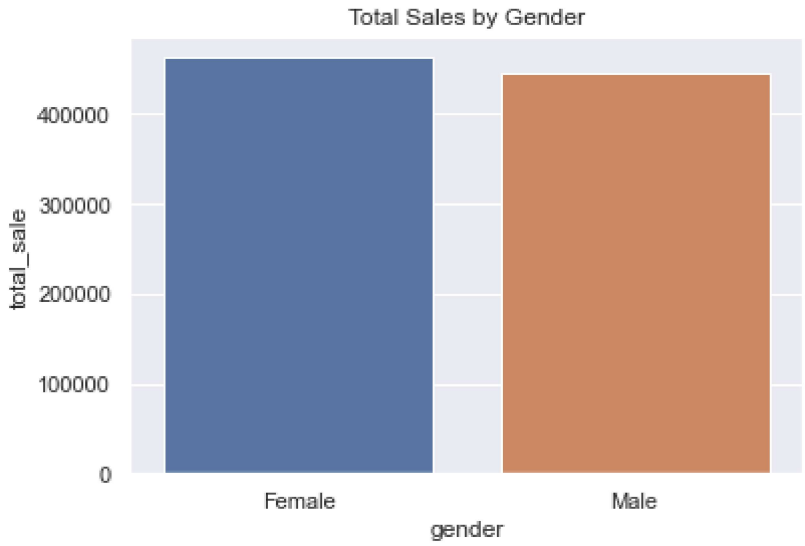


In [49]:
```
df.groupby(['gender'], as_index=False)['total_sale'].sum().sort_values(by=
```

Out[49]:

	gender	total_sale
0	Female	463110.0
1	Male	445120.0

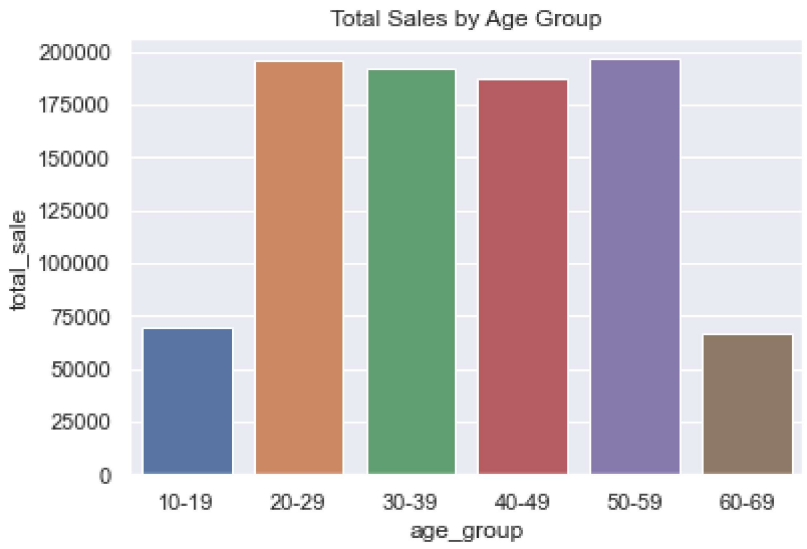
In [58]:
```
# Total sale by gender
sns.barplot(x='gender', y='total_sale', data=df.groupby('gender', as_index
plt.title('Total Sales by Gender')
plt.show()
```



From the above graph, we can see most of the buyers are females

Age

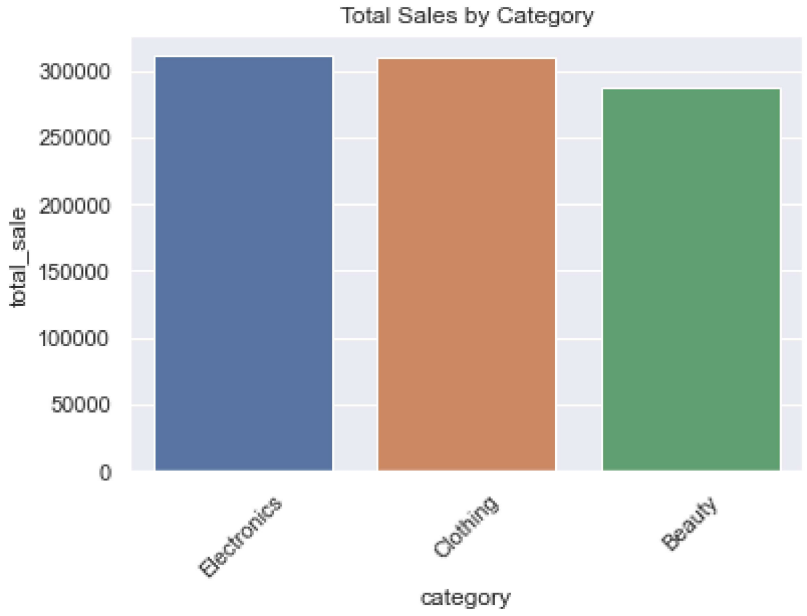
In [59]:
```
df['age_group'] = pd.cut(df['age'], bins=range(10, 80, 10), labels=['10-19
sns.barplot(x='age_group', y='total_sale', data=df.groupby('age_group', as
plt.title('Total Sales by Age Group')
plt.show()
```



50-59 age group is most active in purchases.

Category

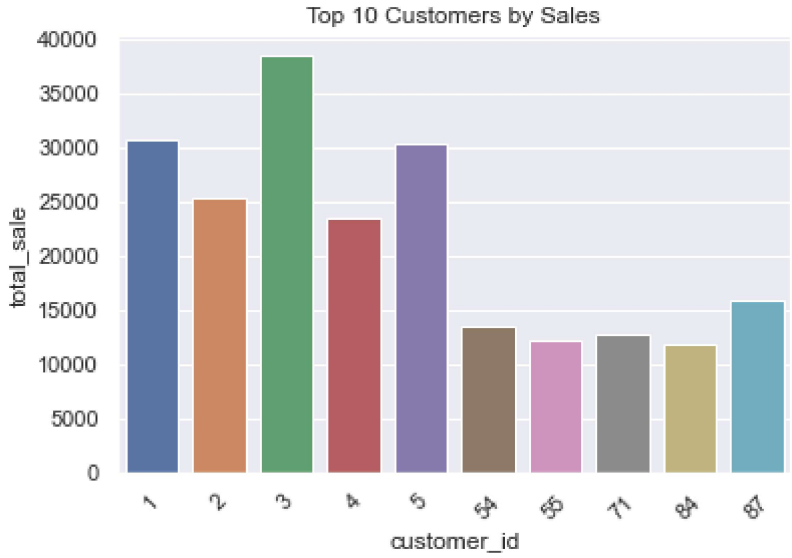
In [57]:
```
# Total sale by category
sns.barplot(x='category', y='total_sale', data=df.groupby('category', as_i
plt.title('Total Sales by Category')
plt.xticks(rotation=45)
plt.show()
```



Electronics and Clothing product types sell the most frequently.

In [65]:
```
# 6. Top Customers

top_customers = df.groupby('customer_id', as_index=False)['total_sale'].su
sns.barplot(x='customer_id', y='total_sale', data=top_customers)
plt.title('Top 10 Customers by Sales')
plt.xticks(rotation=45)
plt.show()
```



Customer_id 3 is the top customer by sales