```
import pandas as pd
In [1]:
       import numpy as np
In [2]: # Load the dataset
       df = pd.read_csv('raw_ecommerce_data.csv')
       # Initial data preview
       print("Initial Dataset:")
       print(df.head())
      Initial Dataset:
                                                Category Quantity Price \
         Order ID Customer Name
                                  Gender
      0 ORD-1000 Customer 0
                                  Male Electronics 6 363.75
                                                              8
      1 ORD-1001 Customer 1
                                   Male Health & Beauty
                                                                  NaN
      2 ORD-1002 Customer 2
                                                              8 452.38
                                                  Books
                                 Female
      3 ORD-1003 Customer 3 Non-Binary
                                                              8 437.44
                                             Electronics
      4 ORD-1004 Customer 4
                                Female Health & Beauty
                                                              1 353.61
             State Payment Method Purchase Date Total
      0 California
                          Cash 2023-01-01 2182.50
      1
          New York
                         PayPal 2023-01-02
                                                 NaN
                         PayPal 2023-01-03 3619.04
      2
          New York
      3
          New York
                          PayPal 2023-01-04 3499.52
          New York Credit Card 2023-01-05 353.61
In [3]: # Check for missing values
       print("\nMissing Values:")
       print(df.isnull().sum())
      Missing Values:
      Order ID
      Customer Name
                      12
      Gender
      Category
      Quantity
                       0
      Price
                      11
      State
                       a
      Payment Method
      Purchase Date
                       0
      Total
                      11
      dtype: int64
In [4]: # Fill missing Customer Names with "Anonymous"
       df['Customer Name'] = df['Customer Name'].fillna('Anonymous')
In [5]: df.head()
```

Out[5]:		Order ID	Customer Name	Gender	Category	Quantity	Price	State	Payment Method	Purchas Dat
	0	ORD- 1000	Customer 0	Male	Electronics	6	363.75	California	Cash	2023-01
	1	ORD- 1001	Customer 1	Male	Health & Beauty	8	NaN	New York	PayPal	2023-01 0
	2	ORD- 1002	Customer 2	Female	Books	8	452.38	New York	PayPal	2023-01 0
	3	ORD- 1003	Customer 3	Non- Binary	Electronics	8	437.44	New York	PayPal	2023-01 0
	4	ORD- 1004	Customer 4	Female	Health & Beauty	1	353.61	New York	Credit Card	2023-01 0
	4									•
In [7]:	df	tail()								
Out[7]:		Order ID	Customer Name	Gender	Category	Quantity	Price	State	Payment Method	Purcha Da
Out[7]:	95			<b>Gender</b> Non-Binary	Health &	<b>Quantity</b>			•	
Out[7]:	95	ID ORD-	<b>Name</b> Customer	Non-	Health & Beauty		348.53	Florida	Method	2023-(
Out[7]:		ORD- 1095 ORD-	Name Customer 95 Customer	Non- Binary Non-	Health & Beauty Home Decor Health &	5	348.53 247.83	Florida	Method  PayPal  Debit	2023-(
Out[7]:	96	ORD- 1095 ORD- 1096 ORD-	Customer 95 Customer 96 Customer	Non- Binary Non- Binary Non-	Health & Beauty Home Decor Health & Beauty	5 6 6	348.53 247.83	Florida Florida California	Method  PayPal  Debit Card	2023-( 2023-(
Out[7]:	96	ORD- 1095 ORD- 1096 ORD- 1097 ORD-	Customer 95 Customer 96 Customer 97 Customer	Non- Binary Non- Binary Non- Binary	Health & Beauty  Home Decor  Health & Beauty  Electronics	5 6 6	348.53 247.83 204.74 275.52	Florida Florida California	PayPal Debit Card Cash Debit	2023-( 2023-( 2023-(
Out[7]:	96 97 98	ORD- 1095 ORD- 1096 ORD- 1097 ORD- 1098 ORD-	Customer 96 Customer 97 Customer 97 Customer 98 Customer	Non- Binary Non- Binary Non- Binary	Health & Beauty  Home Decor  Health & Beauty  Electronics	5 6 6	348.53 247.83 204.74 275.52	Florida Florida California Texas	PayPal  Debit Card  Cash  Debit Card	2023-( 2023-( 2023-( 2023-(

Out[8]:		Order ID	Customer Name	Gender	Category	Quantity	Price	State	Payment Method	Purcha Da	
	0	ORD- 1000	Customer 0	Male	Electronics	6	363.75	California	Cash	2023-(	
	1	ORD- 1001	Customer 1	Male	Health & Beauty	8	NaN	New York	PayPal	2023-(	
	2	ORD- 1002	Customer 2	Female	Books	8	452.38	New York	PayPal	2023-(	
	3	ORD- 1003	Customer 3	Non- Binary	Electronics	8	437.44	New York	PayPal	2023-(	
	4	ORD- 1004	Customer 4	Female	Health & Beauty	1	353.61	New York	Credit Card	2023-(	
	•••			•••							
	95	ORD- 1095	Customer 95	Non- Binary	Health & Beauty	5	348.53	Florida	PayPal	2023-(	
	96	ORD- 1096	Customer 96	Non- Binary	Home Decor	6	247.83	Florida	Debit Card	2023-(	
	97	ORD- 1097	Customer 97	Non- Binary	Health & Beauty	6	204.74	California	Cash	2023-(	
	98	ORD- 1098	Customer 98	Female	Electronics	4	275.52	Texas	Debit Card	2023-(	
	99	ORD- 1099	Customer 99	Female	Clothing	6	20.89	New York	Cash	2023-(	
	100 rows × 10 columns										
	4									<b>&gt;</b>	
In [9]:	<pre># Fill missing Prices with the mean price mean_price = df['Price'].mean() df['Price'] = df['Price'].fillna(mean_price)</pre>										

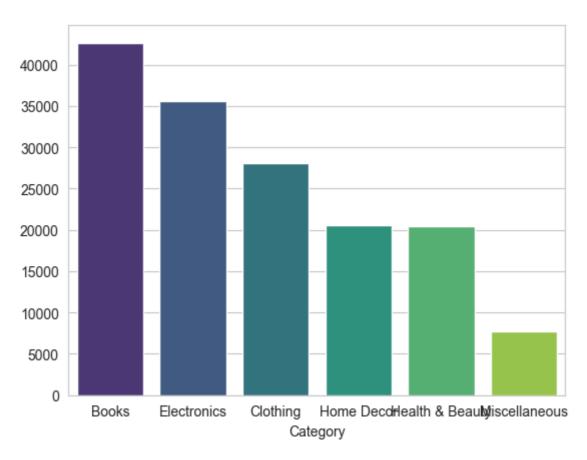
```
In [10]: df
```

Out[10]:		Order ID	Customer Name	Gender	Category	Quantity	Price	State	Payment Method	Pι
	0	ORD- 1000	Customer 0	Male	Electronics	6	363.750000	California	Cash	21
	1	ORD- 1001	Customer 1	Male	Health & Beauty	8	279.048427	New York	PayPal	20
	2	ORD- 1002	Customer 2	Female	Books	8	452.380000	New York	PayPal	21
	3	ORD- 1003	Customer 3	Non- Binary	Electronics	8	437.440000	New York	PayPal	21
	4	ORD- 1004	Customer 4	Female	Health & Beauty	1	353.610000	New York	Credit Card	21
	•••			•••						
	95	ORD- 1095	Customer 95	Non- Binary	Health & Beauty	5	348.530000	Florida	PayPal	21
	96	ORD- 1096	Customer 96	Non- Binary	Home Decor	6	247.830000	Florida	Debit Card	21
	97	ORD- 1097	Customer 97	Non- Binary	Health & Beauty	6	204.740000	California	Cash	21
	98	ORD- 1098	Customer 98	Female	Electronics	4	275.520000	Texas	Debit Card	21
	99	ORD- 1099	Customer 99	Female	Clothing	6	20.890000	New York	Cash	21
	100	rows × 1	10 columns							
	4									•
In [11]:	pri	nt(df.i	snull().su	m())						
Order ID 0 Customer Name 0 Gender 0 Category 0 Quantity 0 Price 0 State 0 Payment Method 0 Purchase Date 0 Total 11 dtype: int64										
In [12]:	<pre>[12]: # Recalculate Total df['Total'] = df['Quantity'] * df['Price']</pre>									
In [13]:	df.	head()								

```
Out[13]:
                                                                                     Payment
              Order
                     Customer
                                                                                               Pui
                                Gender
                                          Category
                                                    Quantity
                                                                    Price
                                                                              State
                 ID
                        Name
                                                                                      Method
              ORD-
                     Customer
                                                                                               202
          0
                                  Male
                                         Electronics
                                                           6 363.750000 California
                                                                                         Cash
               1000
                             0
                                          Health &
              ORD-
                     Customer
                                                                                               202
                                                                                        PayPal
                                  Male
                                                              279.048427
                                                                          New York
               1001
                                            Beauty
              ORD-
                     Customer
                                                                                               202
                                 Female
                                             Books
                                                              452.380000 New York
                                                                                        PayPal
          2
               1002
                             2
              ORD-
                     Customer
                                  Non-
                                                                                               202
          3
                                         Electronics
                                                           8 437.440000
                                                                                        PayPal
                                                                          New York
                                 Binary
               1003
                             3
              ORD-
                     Customer
                                          Health &
                                                                                        Credit
                                                                                               20;
          4
                                 Female
                                                           1 353.610000
                                                                          New York
               1004
                                                                                         Card
                                            Beauty
                             4
                                                                                                # Replace 'Unknown' in Category with 'Miscellaneous'
In [14]:
          df['Category'] = df['Category'].replace('Unknown', 'Miscellaneous')
In [15]:
          # Convert Purchase Date to datetime
          df['Purchase Date'] = pd.to_datetime(df['Purchase Date'])
In [16]:
          df.head()
Out[16]:
             Order
                     Customer
                                                                                     Payment
                                                                                               Pui
                                Gender
                                          Category Quantity
                                                                    Price
                                                                              State
                 ID
                        Name
                                                                                      Method
              ORD-
                     Customer
                                                                                               202
                                                                                         Cash
          0
                                         Electronics
                                                              363.750000 California
                                  Male
               1000
                             0
                                          Health &
              ORD-
                     Customer
                                                                                               202
                                  Male
                                                                                        PayPal
          1
                                                              279.048427
                                                                           New York
               1001
                                            Beauty
              ORD-
                     Customer
                                                                                               202
          2
                                                                                        PayPal
                                 Female
                                             Books
                                                              452.380000
                                                                          New York
               1002
              ORD-
                     Customer
                                                                                                202
                                  Non-
          3
                                         Electronics
                                                              437.440000
                                                                           New York
                                                                                        PayPal
               1003
                             3
                                  Binary
              ORD-
                     Customer
                                          Health &
                                                                                        Credit
                                                                                               202
                                                              353.610000 New York
          4
                                 Female
               1004
                                            Beauty
                                                                                         Card
In [17]:
          # Check cleaned data
          print("\nCleaned Dataset:")
          print(df.info())
```

```
Cleaned Dataset:
       <class 'pandas.core.frame.DataFrame'>
       RangeIndex: 100 entries, 0 to 99
       Data columns (total 10 columns):
                          Non-Null Count Dtype
        # Column
       ---
                           -----
           Order ID 100 non-null
        a
                                         object
        1 Customer Name 100 non-null object
                         100 non-null
        2 Gender
                                         object
           Category
                         100 non-null object
        4 Quantity
                         100 non-null int64
        5 Price
                          100 non-null float64
                          100 non-null object
        6
           State
            Payment Method 100 non-null object
        8  Purchase Date 100 non-null datetime64[ns]
        9
           Total
                          100 non-null
                                         float64
       dtypes: datetime64[ns](1), float64(2), int64(1), object(6)
       memory usage: 7.9+ KB
       None
In [18]: # Save the cleaned dataset
         cleaned_csv_path = 'cleaned_ecommerce_data.csv'
         df.to_csv(cleaned_csv_path, index=False)
         print(f"Cleaned dataset saved: {cleaned_csv_path}")
       Cleaned dataset saved: cleaned_ecommerce_data.csv
         import matplotlib.pyplot as plt
In [23]:
         import seaborn as sns
In [24]: # Set visualization style
         sns.set_style('whitegrid')
         # Top Categories by Revenue
         category_revenue = df.groupby('Category')['Total'].sum().sort_values(ascending=F
         plt.figure(figsize=(10, 6))
Out[24]: <Figure size 1000x600 with 0 Axes>
       <Figure size 1000x600 with 0 Axes>
In [25]: sns.barplot(x=category revenue.index, y=category revenue.values, palette='viridi
       C:\Users\switc\AppData\Local\Temp\ipykernel_17820\3810512292.py:1: FutureWarning:
       Passing `palette` without assigning `hue` is deprecated and will be removed in v
       0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the same effe
         sns.barplot(x=category_revenue.index, y=category_revenue.values, palette='virid
       is')
```

Out[25]: <Axes: xlabel='Category'>



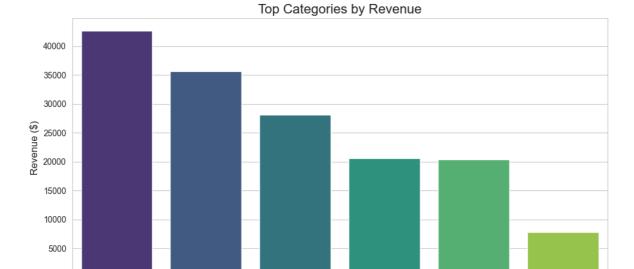
```
In [32]: plt.figure(figsize=(10, 6))
    sns.barplot(x=category_revenue.index, y=category_revenue.values, palette='viridi
    plt.title('Top Categories by Revenue', fontsize=16)
    plt.xlabel('Category', fontsize=12)
    plt.ylabel('Revenue ($)', fontsize=12)
    plt.xticks(rotation=45)
    plt.tight_layout()
    plt.show()
```

C:\Users\switc\AppData\Local\Temp\ipykernel\_17820\3870076080.py:2: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v 0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the same effect.

sns.barplot(x=category\_revenue.index, y=category\_revenue.values, palette='virid
is')

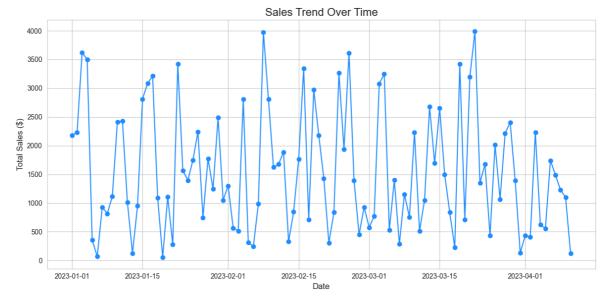
0



Category

```
In [28]: # Sales Trend Over Time
    sales_trend = df.groupby('Purchase Date')['Total'].sum()

plt.figure(figsize=(12, 6))
    plt.plot(sales_trend.index, sales_trend.values, marker='o', color='dodgerblue')
    plt.title('Sales Trend Over Time', fontsize=16)
    plt.xlabel('Date', fontsize=12)
    plt.ylabel('Total Sales ($)', fontsize=12)
    plt.grid(True)
    plt.tight_layout()
    plt.savefig('sales_trend.png')
    plt.show()
```



```
In [41]: import plotly.express as px
In [40]: # Payment Method Distribution
    payment_counts = df['Payment Method'].value_counts()
    fig = px.pie(
        values=payment_counts.values,
```

```
names=payment_counts.index,
  title='Payment Method Distribution',
  color_discrete_sequence=px.colors.sequential.RdBu
)
fig.show()
```

## Payment Method Distribution



In [ ]: