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

sns.set(style="whitegrid")
plt.rcParams["figure.figsize"] = (10, 5)

# Load dataset
df = pd.read_csv(r"C:\Users\Sunayana Panigrahi\Downloads\archive (2)\
train.csv")
df
```

```
      Row ID        Order ID  Order Date   Ship Date       Ship
Mode  \
0          1  CA-2017-152156  08/11/2017  11/11/2017    Second Class

1          2  CA-2017-152156  08/11/2017  11/11/2017    Second Class

2          3  CA-2017-138688  12/06/2017  16/06/2017    Second Class

3          4  US-2016-108966  11/10/2016  18/10/2016  Standard Class

4          5  US-2016-108966  11/10/2016  18/10/2016  Standard Class

...      ...             ...         ...         ...             ...

9795    9796  CA-2017-125920  21/05/2017  28/05/2017  Standard Class

9796    9797  CA-2016-128608  12/01/2016  17/01/2016  Standard Class

9797    9798  CA-2016-128608  12/01/2016  17/01/2016  Standard Class

9798    9799  CA-2016-128608  12/01/2016  17/01/2016  Standard Class

9799    9800  CA-2016-128608  12/01/2016  17/01/2016  Standard Class


     Customer ID     Customer Name    Segment        Country
City  \
0       CG-12520       Claire Gute   Consumer  United States
Henderson
1       CG-12520       Claire Gute   Consumer  United States
Henderson
2       DV-13045   Darrin Van Huff  Corporate  United States      Los
Angeles
3       SO-20335    Sean O'Donnell   Consumer  United States  Fort
Lauderdale
4       SO-20335    Sean O'Donnell   Consumer  United States  Fort
Lauderdale
...          ...               ...         ...             ...
...
```

```
9795    SH-19975     Sally Hughsby  Corporate  United States
Chicago
9796    CS-12490  Cindy Schnelling  Corporate  United States
Toledo
9797    CS-12490  Cindy Schnelling  Corporate  United States
Toledo
9798    CS-12490  Cindy Schnelling  Corporate  United States
Toledo
9799    CS-12490  Cindy Schnelling  Corporate  United States
Toledo

           State  Postal Code   Region       Product ID
Category  \
0       Kentucky      42420.0    South  FUR-BO-10001798
Furniture
1       Kentucky      42420.0    South  FUR-CH-10000454
Furniture
2     California      90036.0     West  OFF-LA-10000240  Office
Supplies
3        Florida      33311.0    South  FUR-TA-10000577
Furniture
4        Florida      33311.0    South  OFF-ST-10000760  Office
Supplies
...          ...          ...      ...              ...               .
..
9795    Illinois      60610.0  Central  OFF-BI-10003429  Office
Supplies
9796        Ohio      43615.0     East  OFF-AR-10001374  Office
Supplies
9797        Ohio      43615.0     East  TEC-PH-10004977
Technology
9798        Ohio      43615.0     East  TEC-PH-10000912
Technology
9799        Ohio      43615.0     East  TEC-AC-10000487
Technology

     Sub-Category                                         Product Name
Sales
0       Bookcases                  Bush Somerset Collection Bookcase
261.9600
1         Chairs  Hon Deluxe Fabric Upholstered Stacking Chairs,...
731.9400
2         Labels  Self-Adhesive Address Labels for Typewriters b...
14.6200
3         Tables      Bretford CR4500 Series Slim Rectangular Table
957.5775
4        Storage                     Eldon Fold 'N Roll Cart System
22.3680
...          ...                                                ...
```

```
...
9795      Binders  Cardinal HOLDit! Binder Insert Strips,Extra St...
3.7980
9796          Art           BIC Brite Liner Highlighters, Chisel Tip
10.3680
9797       Phones                                          GE 30524EE4
235.1880
9798       Phones           Anker 24W Portable Micro USB Car Charger
26.3760
9799  Accessories                SanDisk Cruzer 4 GB USB Flash Drive
10.3840

[9800 rows x 18 columns]
```

```
# Drop duplicates
df.drop_duplicates(inplace=True)

# Clean the dataset
df.columns = df.columns.str.strip().str.lower().str.replace(' ', '_')
df['order_date'] = pd.to_datetime(df['order_date'], dayfirst=True)
df['ship_date'] = pd.to_datetime(df['ship_date'], dayfirst=True)

# Check for missing values
df.isnull().sum()
```

```
Row ID            0
Order ID          0
Order Date        0
Ship Date         0
Ship Mode         0
Customer ID       0
Customer Name     0
Segment           0
Country           0
City              0
State             0
Postal Code      11
Region            0
Product ID        0
Category          0
Sub-Category      0
Product Name      0
Sales             0
dtype: int64
```

```
df.columns
```

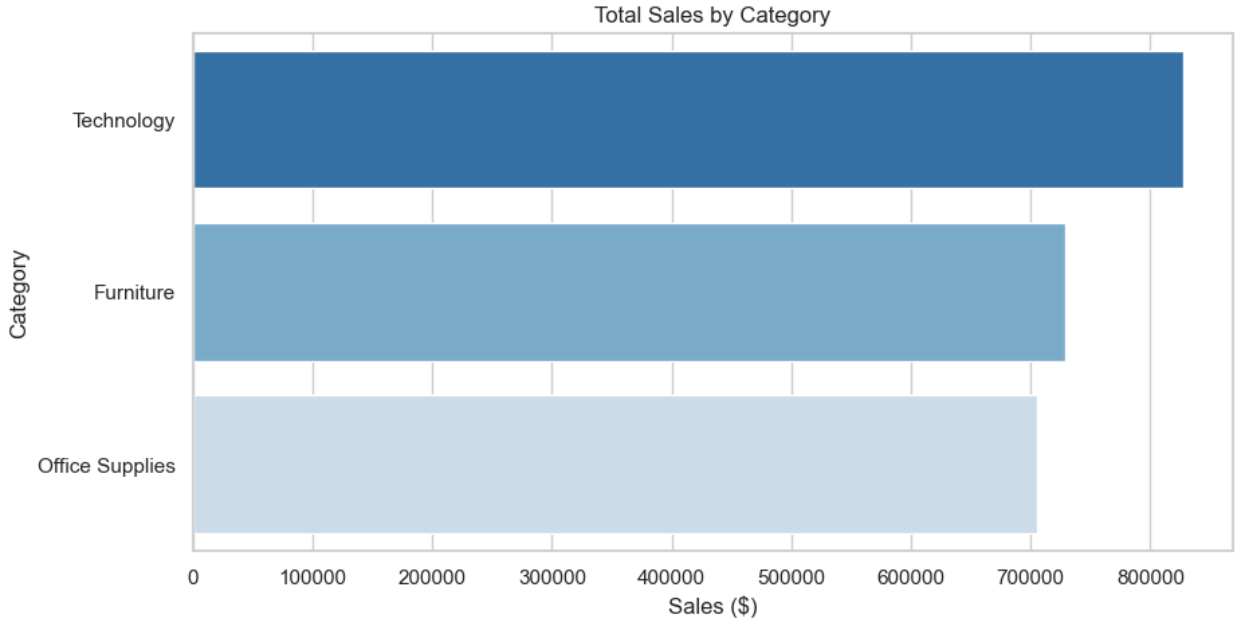
```
Index(['Row ID', 'Order ID', 'Order Date', 'Ship Date', 'Ship Mode',
       'Customer ID', 'Customer Name', 'Segment', 'Country', 'City',
'State',
       'Postal Code', 'Region', 'Product ID', 'Category', 'Sub-
```

```
Category',
       'Product Name', 'Sales'],
      dtype='object')

import warnings
warnings.filterwarnings('ignore')

#sales by category
category_sales = df.groupby('Category')
['Sales'].sum().sort_values(ascending=False)

sns.barplot(x=category_sales.values, y=category_sales.index,
palette='Blues_r')
plt.title("Total Sales by Category")
plt.xlabel("Sales ($)")
plt.ylabel("Category")
plt.show()
```



Sales by Category – Key Insights

Technology leads in sales – high demand and high-value items.

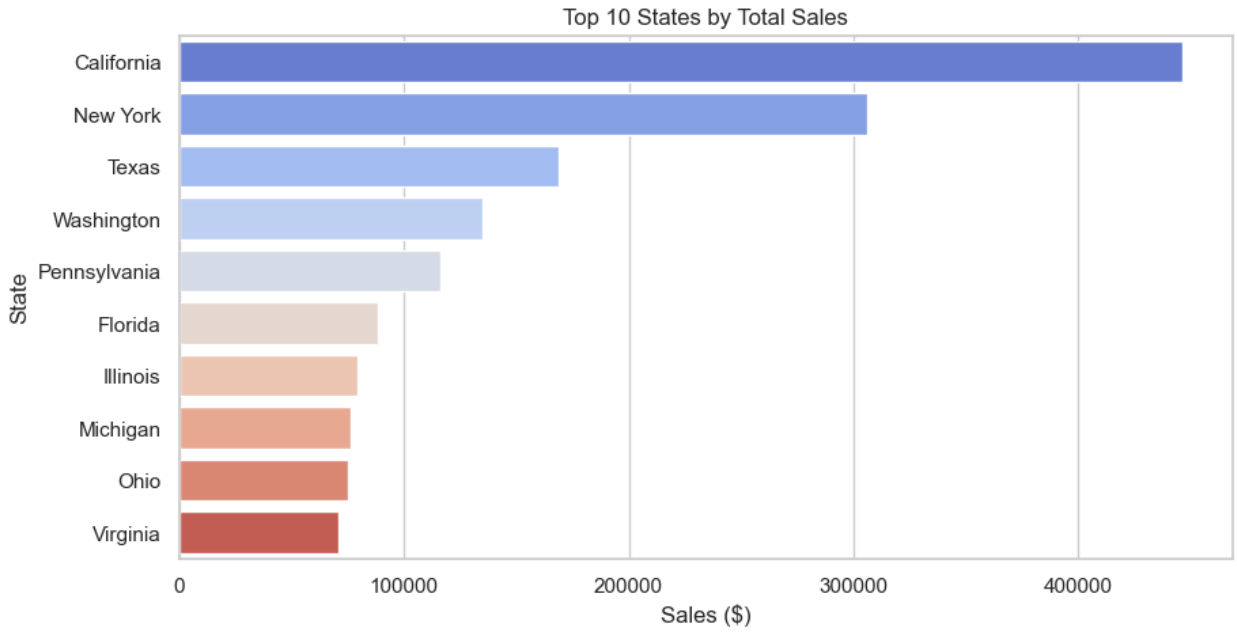
Furniture has moderate sales – room for growth.

Office Supplies has lowest sales – frequent but low-value purchases.

```
#"Top 10 States by Total Sales
top_states = df.groupby('State')
['Sales'].sum().sort_values(ascending=False).head(10)

sns.barplot(x=top_states.values, y=top_states.index,
```

```
palette='coolwarm')
plt.title("Top 10 States by Total Sales")
plt.xlabel("Sales ($)")
plt.ylabel("State")
plt.show()
```



Top 10 States by Sales – Key Insights

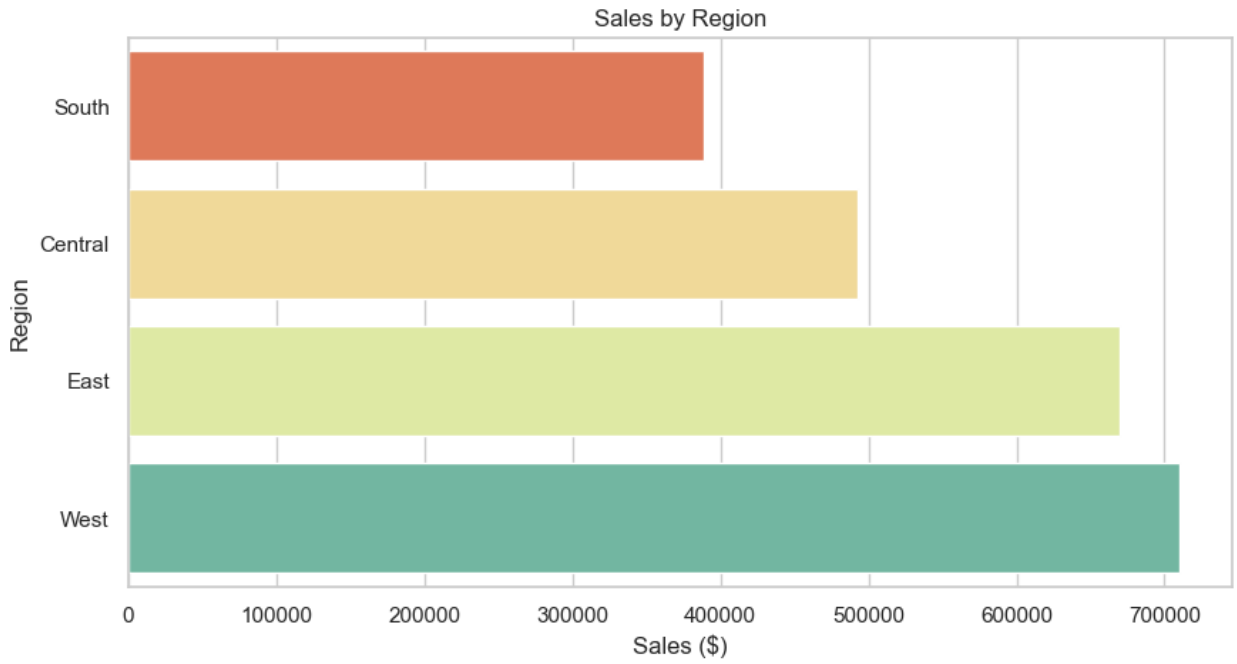
California dominates sales by a large margin — it's the primary revenue driver.

New York, Texas, and Washington follow with strong performance.

Other top states (e.g., Michigan, Pennsylvania) contribute steadily but with lower volume.

```
#Sales by Region
region_sales = df.groupby('Region')['Sales'].sum().sort_values()

sns.barplot(x=region_sales.values, y=region_sales.index,
palette='Spectral')
plt.title("Sales by Region")
plt.xlabel("Sales ($)")
plt.ylabel("Region")
plt.show()
```



Sales by Region – Key Insights

West region leads in total sales, followed by East.

Central and South regions have lower sales comparatively.

Strengthen marketing in Central and South to balance regional performance.

```
# clean dataset
df.to_csv("cleaned_superstore.csv", index=False)
```

Superstore Sales Analysis Summary

Key Insights:

- *Technology* is the top-performing category by sales.
- *California* leads among all states in total sales.
- Sales increased steadily from *2014 to 2017*.
- *West and East* regions contribute most to revenue.

Strategic Suggestions:

- Boost sales in underperforming regions with targeted campaigns.
- Expand high-selling categories like *Technology* into new states.
- Maintain sales growth momentum with seasonal promotions and offers.

```
```