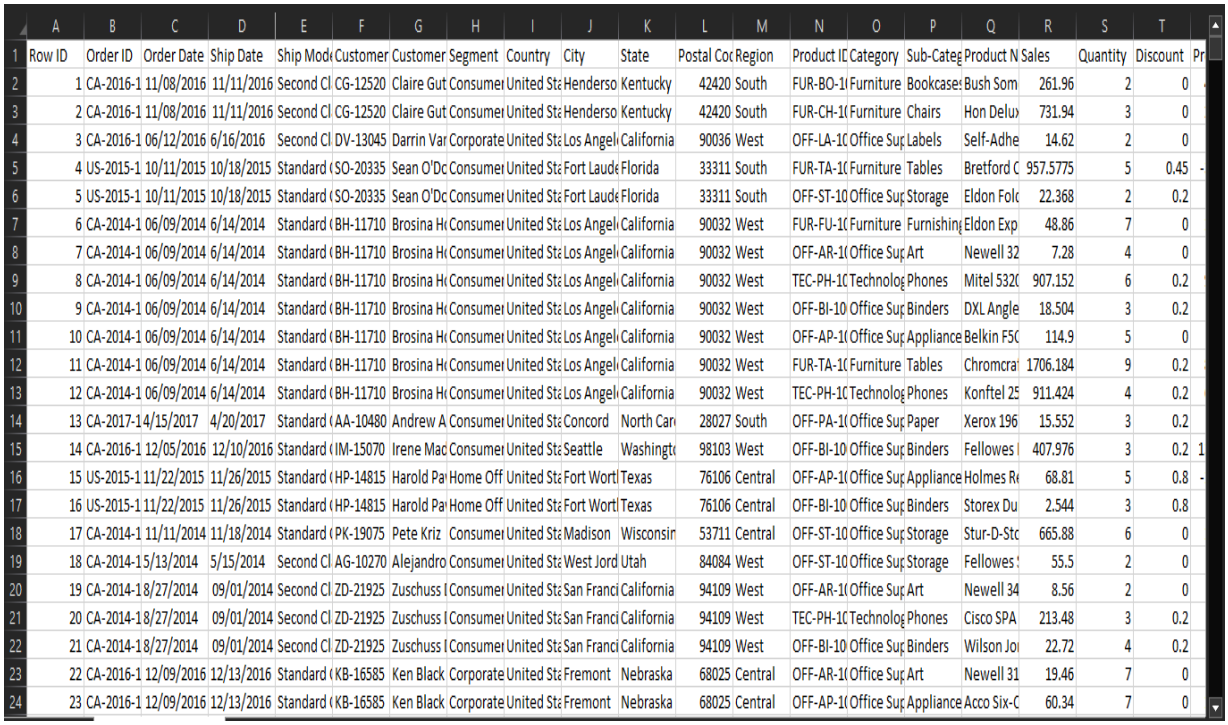
Exploratory Data Analysis of Superstore Sales Data

Objective:

Analyze the Superstore dataset to uncover insights on:
Regional performance
Customer segments
Product categories & sub-categories
Discount policies
Shipping efficiency
…and recommend actionable strategies.

Dataset Description:

- **Name:** Sample - Superstore.csv
- **Records:** ~10,000 transactions
- **Features:**
    - Order & Ship Dates
    - Customer Name, Segment & Region
    - Product Category & Sub-Category
    - Sales, Quantity, Discount, Profit
- **Period Covered:** Historical transaction data



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	
1	Row ID	Order ID	Order Date	Ship Date	Ship Mod	Customer	Customer	Segment	Country	City	State	Postal Co	Region	Product ID	Category	Sub-Categ	Product N	Sales	Quantity	Discount	Pr
2	1	CA-2016-1	11/08/2016	11/11/2016	Second Cl	CG-12520	Claire Gut	Consumer	United Sta	Henderso	Kentucky	42420	South	FUR-BO-1(	Furniture	Bookcases	Bush Som	261.96	2	0	
3	2	CA-2016-1	11/08/2016	11/11/2016	Second Cl	CG-12520	Claire Gut	Consumer	United Sta	Henderso	Kentucky	42420	South	FUR-CH-1(	Furniture	Chairs	Hon Delux	731.94	3	0	
4	3	CA-2016-1	06/12/2016	6/16/2016	Second Cl	DV-13045	Darrin Van	Corporate	United Sta	Los Angel	California	90036	West	OFF-LA-1(	Office Sup	Labels	Self-Adhe	14.62	2	0	
5	4	US-2015-1	10/11/2015	10/18/2015	Standard (	SO-20335	Sean O'Dc	Consumer	United Sta	Fort Laud	Florida	33311	South	FUR-TA-1(	Furniture	Tables	Bretford C	957.5775	5	0.45	-
6	5	US-2015-1	10/11/2015	10/18/2015	Standard (	SO-20335	Sean O'Dc	Consumer	United Sta	Fort Laud	Florida	33311	South	OFF-ST-10	Office Sup	Storage	Eldon Fol	22.368	2	0.2	
7	6	CA-2014-1	06/09/2014	6/14/2014	Standard (	BH-11710	Brosina H	Consumer	United Sta	Los Angel	California	90032	West	FUR-FU-1(	Furniture	Furnishing	Eldon Exp	48.86	7	0	
8	7	CA-2014-1	06/09/2014	6/14/2014	Standard (	BH-11710	Brosina H	Consumer	United Sta	Los Angel	California	90032	West	OFF-AR-1(	Office Sup	Art	Newell 32	7.28	4	0	
9	8	CA-2014-1	06/09/2014	6/14/2014	Standard (	BH-11710	Brosina H	Consumer	United Sta	Los Angel	California	90032	West	TEC-PH-1(	Technolog	Phones	Mitel 532(	907.152	6	0.2	
10	9	CA-2014-1	06/09/2014	6/14/2014	Standard (	BH-11710	Brosina H	Consumer	United Sta	Los Angel	California	90032	West	OFF-BI-10	Office Sup	Binders	DXL Angle	18.504	3	0.2	
11	10	CA-2014-1	06/09/2014	6/14/2014	Standard (	BH-11710	Brosina H	Consumer	United Sta	Los Angel	California	90032	West	OFF-AP-1(	Office Sup	Appliance	Belkin F5(	114.9	5	0	
12	11	CA-2014-1	06/09/2014	6/14/2014	Standard (	BH-11710	Brosina H	Consumer	United Sta	Los Angel	California	90032	West	FUR-TA-1(	Furniture	Tables	Chromcra	1706.184	9	0.2	
13	12	CA-2014-1	06/09/2014	6/14/2014	Standard (	BH-11710	Brosina H	Consumer	United Sta	Los Angel	California	90032	West	TEC-PH-1(	Technolog	Phones	Konftel 25	911.424	4	0.2	
14	13	CA-2017-1	4/15/2017	4/20/2017	Standard (	AA-10480	Andrew A	Consumer	United Sta	Concord	North Car	28027	South	OFF-PA-1(	Office Sup	Paper	Xerox 196	15.552	3	0.2	
15	14	CA-2016-1	12/05/2016	12/10/2016	Standard (	IM-15070	Irene Mac	Consumer	United Sta	Seattle	Washingt	98103	West	OFF-BI-10	Office Sup	Binders	Fellowes	407.976	3	0.2	1
16	15	US-2015-1	11/22/2015	11/26/2015	Standard (	HP-14815	Harold Pa	Home Off	United Sta	Fort Wort	Texas	76106	Central	OFF-AP-1(	Office Sup	Appliance	Holmes Re	68.81	5	0.8	-
17	16	US-2015-1	11/22/2015	11/26/2015	Standard (	HP-14815	Harold Pa	Home Off	United Sta	Fort Wort	Texas	76106	Central	OFF-BI-10	Office Sup	Binders	Storex Du	2.544	3	0.8	
18	17	CA-2014-1	11/11/2014	11/18/2014	Standard (	PK-19075	Pete Kriz	Consumer	United Sta	Madison	Wisconsir	53711	Central	OFF-ST-10	Office Sup	Storage	Stur-D-Sto	665.88	6	0	
19	18	CA-2014-1	5/13/2014	5/15/2014	Second Cl	AG-10270	Alejandro	Consumer	United Sta	West Jord	Utah	84084	West	OFF-ST-10	Office Sup	Storage	Fellowes	55.5	2	0	
20	19	CA-2014-1	8/27/2014	09/01/2014	Second Cl	ZD-21925	Zuschuss I	Consumer	United Sta	San Franci	California	94109	West	OFF-AR-1(	Office Sup	Art	Newell 34	8.56	2	0	
21	20	CA-2014-1	8/27/2014	09/01/2014	Second Cl	ZD-21925	Zuschuss I	Consumer	United Sta	San Franci	California	94109	West	TEC-PH-1(	Technolog	Phones	Cisco SPA	213.48	3	0.2	
22	21	CA-2014-1	8/27/2014	09/01/2014	Second Cl	ZD-21925	Zuschuss I	Consumer	United Sta	San Franci	California	94109	West	OFF-BI-10	Office Sup	Binders	Wilson Jo	22.72	4	0.2	
23	22	CA-2016-1	12/09/2016	12/13/2016	Standard (	KB-16585	Ken Black	Corporate	United Sta	Fremont	Nebraska	68025	Central	OFF-AR-1(	Office Sup	Art	Newell 31	19.46	7	0	
24	23	CA-2016-1	12/09/2016	12/13/2016	Standard (	KB-16585	Ken Black	Corporate	United Sta	Fremont	Nebraska	68025	Central	OFF-AP-1(	Office Sup	Appliance	Acco Six-C	60.34	7	0	

Business Questions:

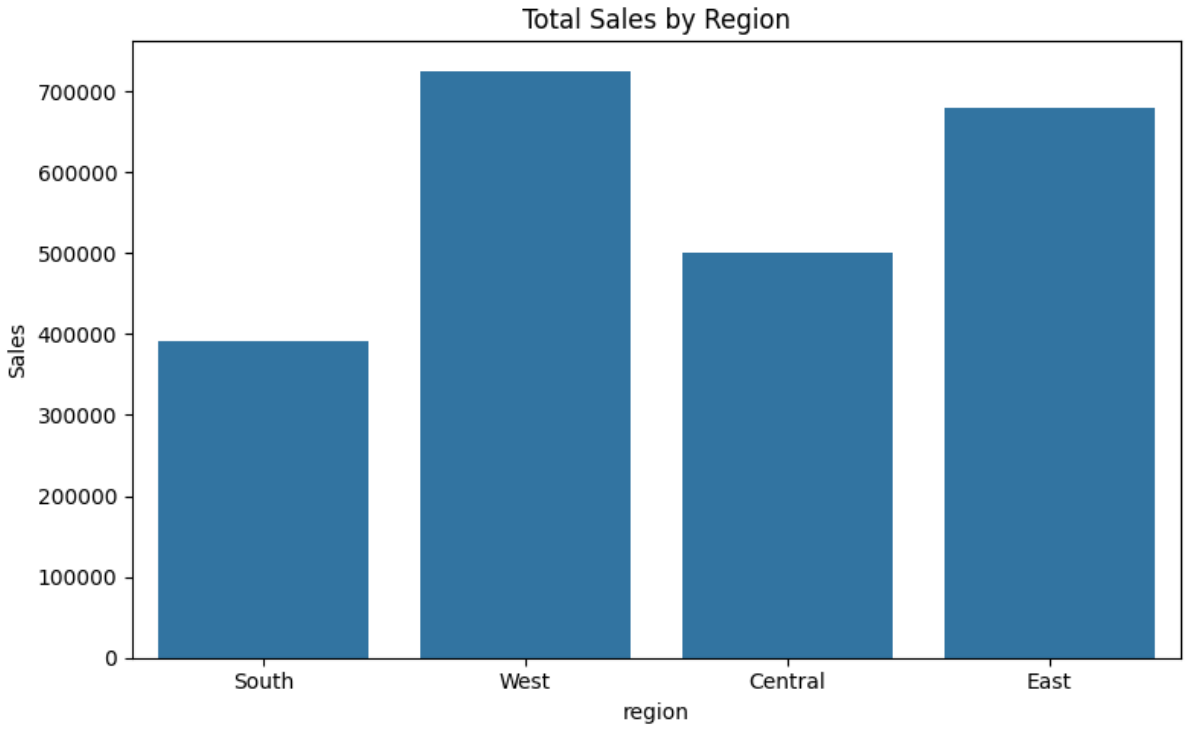
Which regions generate the highest sales & profit?

```
region_summary = df.groupby('region')[['sales', 'profit']].sum().sort_values(by='sales', ascending=False)
print(region_summary)

plt.figure(figsize=(8,5))
sns.barplot(x='region', y='sales', data=df, estimator=sum, ci=None)
plt.title('Total Sales by Region')
plt.ylabel('Sales')
plt.tight_layout()
plt.savefig('sales_by_region.png')
plt.show()

plt.figure(figsize=(8,5))
sns.barplot(x='region', y='profit', data=df, estimator=sum, ci=None)
plt.title('Total Profit by Region')
plt.ylabel('Profit')
plt.tight_layout()
plt.savefig('profit_by_region.png')
plt.show()
```

```
              sales       profit
region
West     725457.8245  108418.4489
East     678781.2400   91522.7800
Central  501239.8908   39706.3625
South    391721.9050   46749.4303
/tmp/ipython-input-87-388932715.py:5: FutureWarning:
```



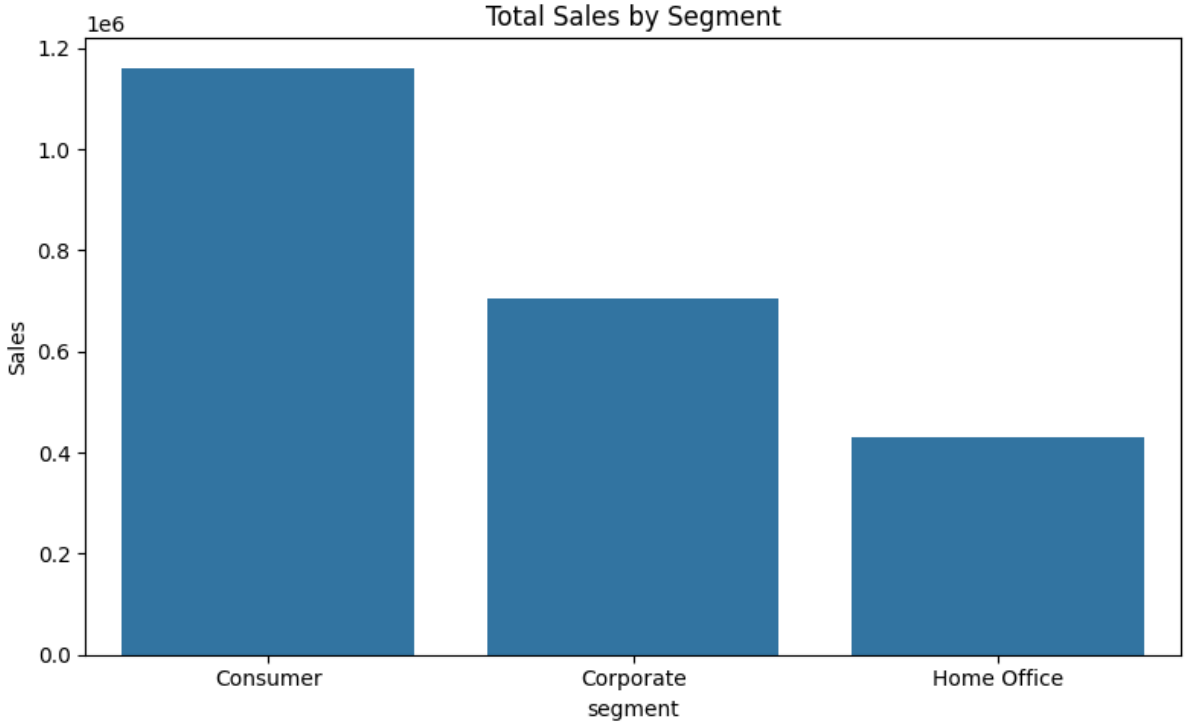
Which customer segments contribute most to profit?

```
segment_summary = df.groupby('segment')[['sales', 'profit']].sum().sort_values(by='sales', ascending=False)
print(segment_summary)

plt.figure(figsize=(8,5))
sns.barplot(x='segment', y='sales', data=df, estimator=sum, ci=None)
plt.title('Total Sales by Segment')
plt.ylabel('Sales')
plt.tight_layout()
plt.savefig('sales_by_segment.png')
plt.show()

plt.figure(figsize=(8,5))
sns.barplot(x='segment', y='profit', data=df, estimator=sum, ci=None)
plt.title('Total Profit by Segment')
plt.ylabel('Profit')
plt.tight_layout()
plt.savefig('profit_by_segment.png')
plt.show()
```

```
                  sales       profit
segment
Consumer     1.161401e+06  134119.2092
Corporate    7.061464e+05   91979.1340
Home Office  4.296531e+05   60298.6785
/tmp/ipython-input-88-1150673367.py:5: FutureWarning:
```

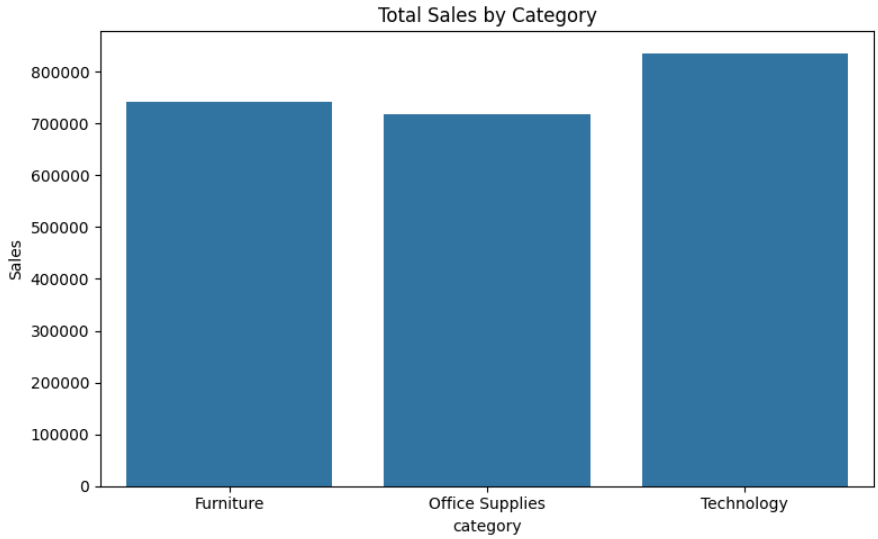


Which product categories and sub-categories perform best?

```
plt.figure(figsize=(8,5))
sns.barplot(x='category', y='sales', data=
plt.title('Total Sales by Category')
plt.ylabel('Sales')
plt.tight_layout()
plt.savefig('sales_by_category.png')
plt.show()

plt.figure(figsize=(12,6))
sns.barplot(x='sub-category', y='sales', data=df, estimator=sum, ci=None)
plt.title('Total Sales by Sub-Category')
plt.ylabel('Sales')
plt.xticks(rotation=45)
plt.tight_layout()
plt.savefig('sales_by_subcategory.png')
plt.show()
```

```
                     sales       profit
category
Furniture        741999.7953   18451.2728
Office Supplies  719047.0320  122490.8008
Technology       836154.0330  145454.9481
                    sales      profit
sub-category
Phones         330007.0540  44515.7306
```

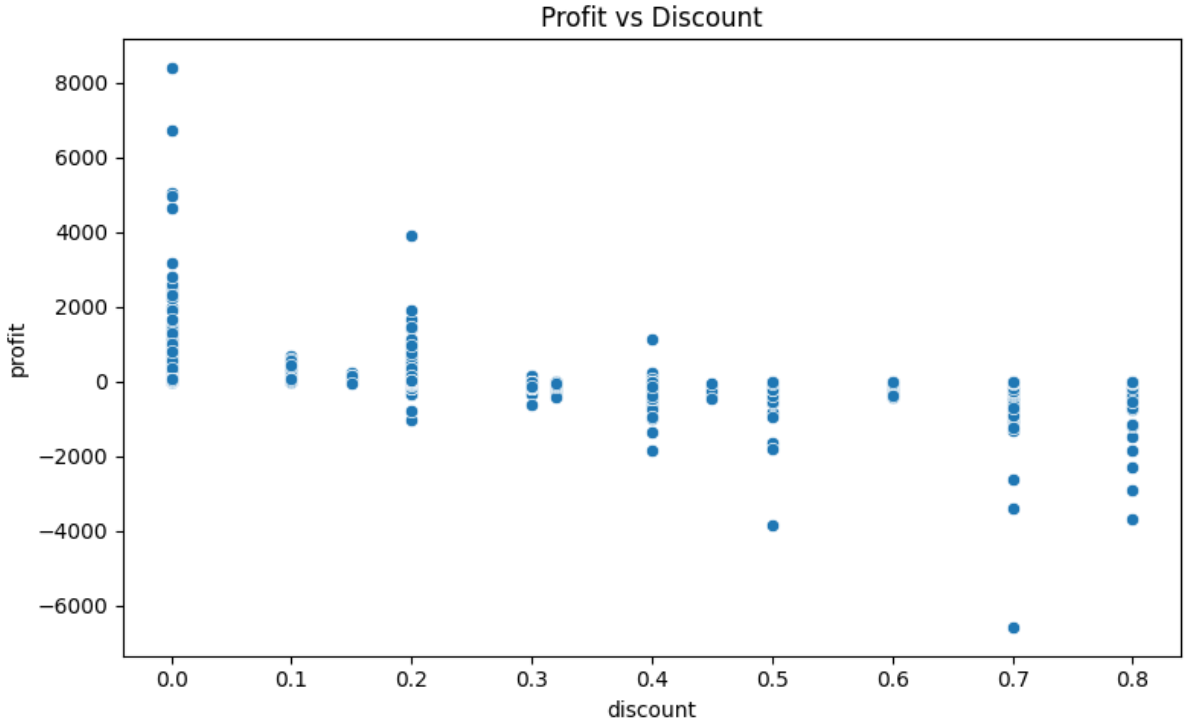


How do discounts impact profitability?

```
print(df[['discount', 'profit']].corr())

plt.figure(figsize=(8,5))
sns.scatterplot(x='discount', y='profit', data=df)
plt.title('Profit vs Discount')
plt.tight_layout()
plt.savefig('profit_vs_discount.png')
plt.show()
```

```
            discount    profit
discount    1.000000 -0.219487
profit     -0.219487  1.000000
```

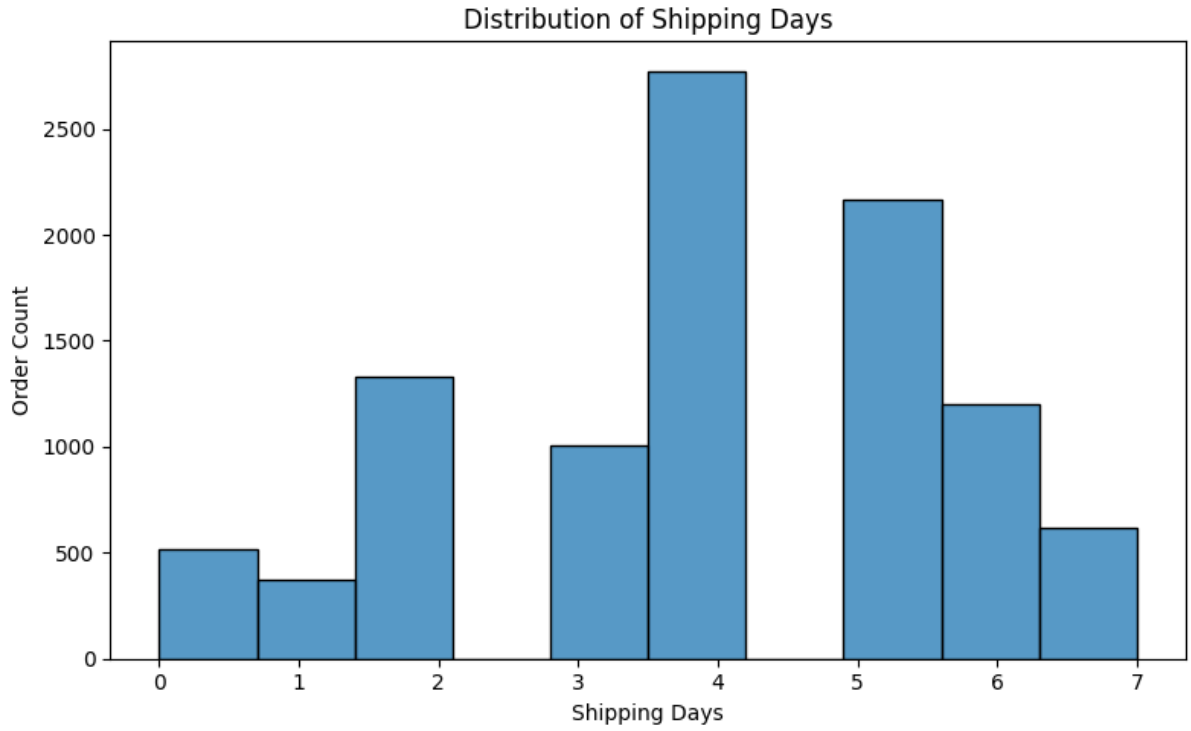


How efficient is the shipping process?

```
print("Average Shipping Days:", df['shipping_days'].mean())

plt.figure(figsize=(8,5))
sns.histplot(df['shipping_days'], bins=10, kde=False)
plt.title('Distribution of Shipping Days')
plt.xlabel('Shipping Days')
plt.ylabel('Order Count')
plt.tight_layout()
plt.savefig('shipping_days_distribution.png')
plt.show()
```

```
Average Shipping Days: 3.958174904942966
```



Data Cleaning:

What it does:

Removes duplicates
Drops rows with missing values
Standardizes column names to snake_case
Converts order_date & ship_date to datetime
Computes shipping_days
Drops irrelevant columns like row_id & postal_code
Saves a clean version as Superstore_Cleaned.csv

Steps:

- **Removed missing & duplicate records**
  - Checked for and eliminated rows with missing critical fields.
  - Dropped exact duplicate rows to avoid over-representing any transaction.
- **Converted date columns**
  - Converted Order Date and Ship Date columns from text to proper datetime format.
  - Enabled calculation of derived fields like shipping duration.
- **Standardized column names**
  - Renamed columns to lowercase.
  - Replaced spaces with underscores for easier code handling.

- Example: Order Date → order_date, Ship Date → ship_date.

- **Removed irrelevant columns**
  - Dropped columns that are not meaningful for business analysis, such as:
    - Row ID
    - Postal Code

```
import pandas as pd

# Load dataset
df = pd.read_csv("Sample - Superstore.csv", encoding='latin1')

# 📋 Inspect data
print(df.shape)
print(df.info())
print(df.head())

# 🧹 Remove duplicates
df.drop_duplicates(inplace=True)

# 🧹 Remove rows with missing values in critical columns
# (adjust columns as needed; here assuming none are critical and missing)
df.dropna(inplace=True)

# 🔤 Standardize column names
df.columns = df.columns.str.strip().str.lower().str.replace(' ', '_')

# 🗓️ Convert date columns
df['order_date'] = pd.to_datetime(df['order_date'])
df['ship_date'] = pd.to_datetime(df['ship_date'])
```

```
 ---  ------         --------------  -----
  0   order_id       9994 non-null   object
  1   order_date     9994 non-null   datetime64[ns]
  2   ship_date      9994 non-null   datetime64[ns]
  3   ship_mode      9994 non-null   object
  4   customer_id    9994 non-null   object
  5   customer_name  9994 non-null   object
  6   segment        9994 non-null   object
  7   country        9994 non-null   object
  8   city           9994 non-null   object
  9   state          9994 non-null   object
  10  region         9994 non-null   object
  11  product_id     9994 non-null   object
  12  category       9994 non-null   object
  13  sub-category   9994 non-null   object
  14  product_name   9994 non-null   object
  15  sales          9994 non-null   float64
  16  quantity       9994 non-null   int64
  17  discount       9994 non-null   float64
  18  profit         9994 non-null   float64
  19  shipping_days  9994 non-null   int64
dtypes: datetime64[ns](2), float64(3), int64(2), object(13)
```

Loaded the data → Cleaned it → Removed irrelevant columns → Converted dates → Standardized names → Added shipping_days.

dataset has:
9,994 rows × 20 columns

order_date & ship_date are proper datetimes
shipping_days is calculated
row_id & postal_code removed
column names are clean (lowercase + underscores)

Superstore EDA — Results

Region-wise Sales & Profit

Region	Sales	Profit
West	725,457.82	108,418.45
East	678,781.24	91,522.78
Central	501,239.89	39,706.36
South	391,721.90	46,749.43

📌 *Insight: West region leads both in sales and profit; South region is weakest.*

Segment-wise Sales & Profit

Segment	Sales	Profit
Consumer	1,161,401.00	134,119.21
Corporate	706,146.40	91,979.13
Home Office	429,653.10	60,298.68

📌 *Insight: Consumer segment contributes the most to sales and profit; Home Office is smallest.*

Category-wise Sales & Profit

Category	Sales	Profit
Furniture	741,999.79	18,451.27
Office Supplies	719,047.03	122,490.80
Technology	836,154.03	145,454.95

📌 *Insight: Technology is the most profitable; Furniture has high sales but very low profit.*

Top 5 Sub-Categories by Sales

Sub-Category	Sales	Profit
Phones	330,007.05	44,515.73
Chairs	328,449.10	26,590.17
Storage	223,843.61	21,278.83
Tables	206,965.53	-17,725.48
Binders	203,412.73	30,221.76

📌 *Insight: Phones & Chairs dominate sales; Tables incur losses despite high sales.*

Discount vs Profit

Correlation between **Discount** and **Profit**:
-0.219 → Higher discounts are associated with lower profits.

📌 *Insight: Keep discounts below 30% where possible to protect margins.*

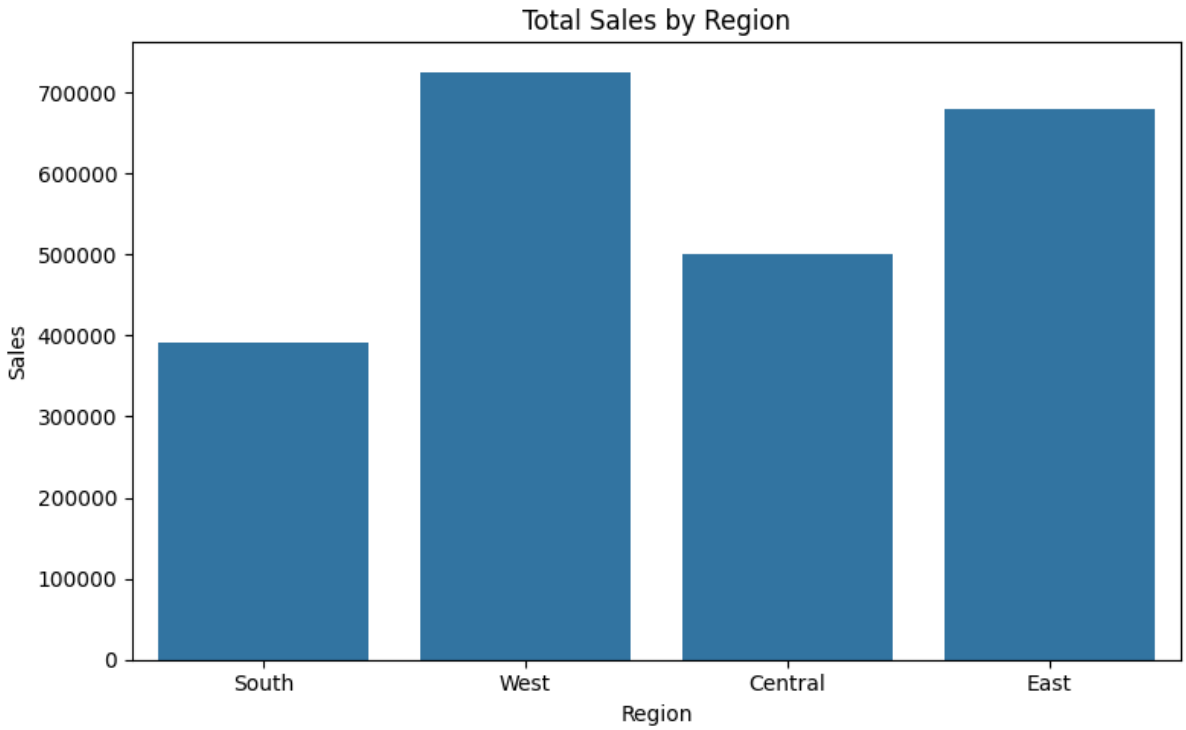
Shipping Efficiency

Average shipping time: **~3.96 days**

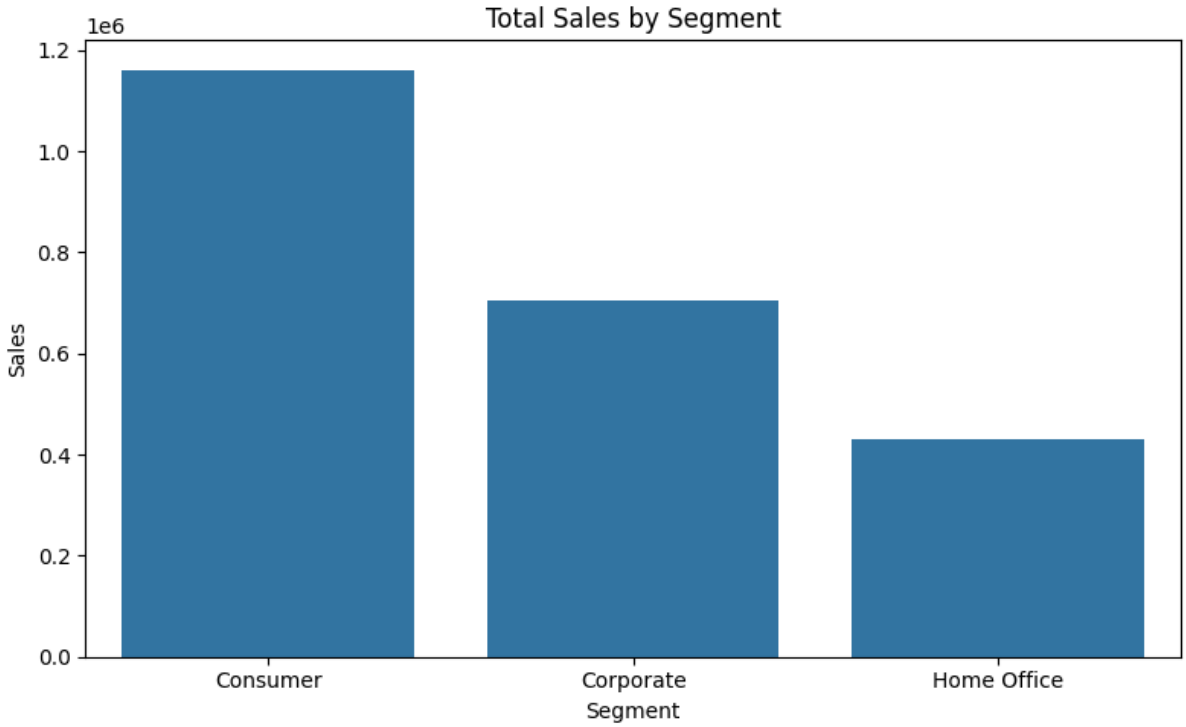
📌 *Insight: Shipping performance is good and should be maintained.*

Visualizations:

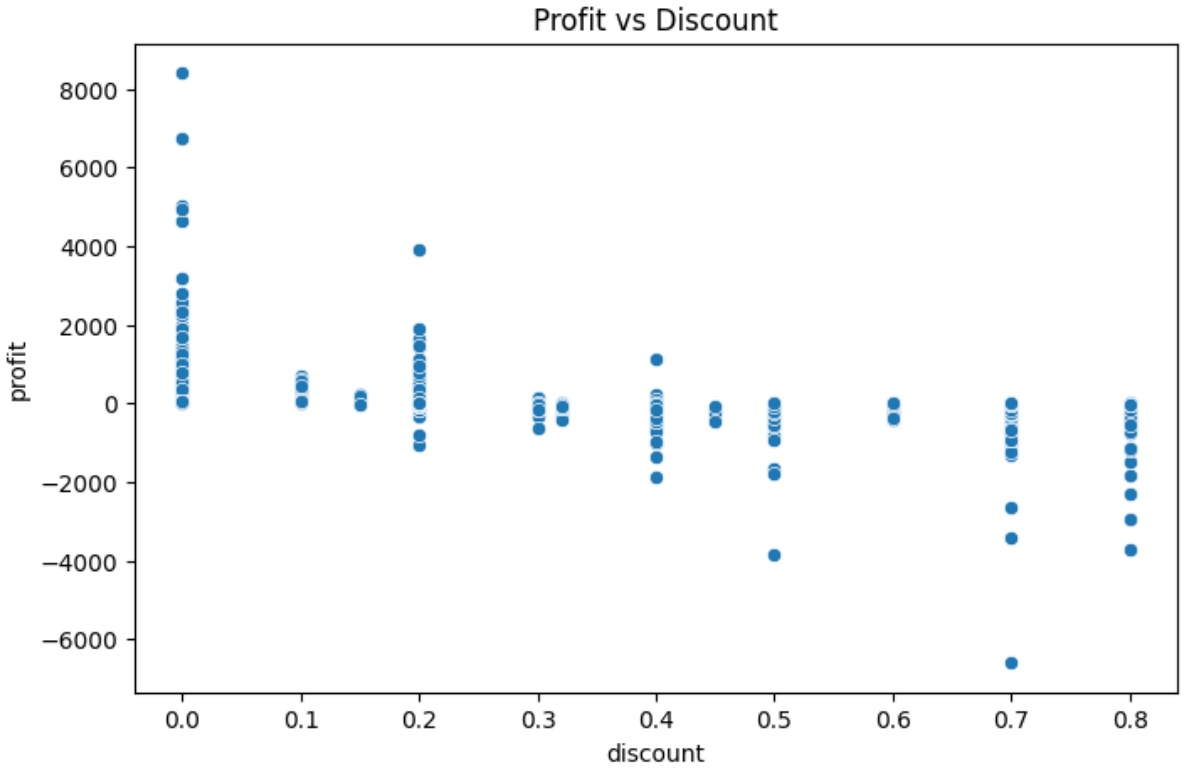
- Barplots:
  - Sales & Profit by Region



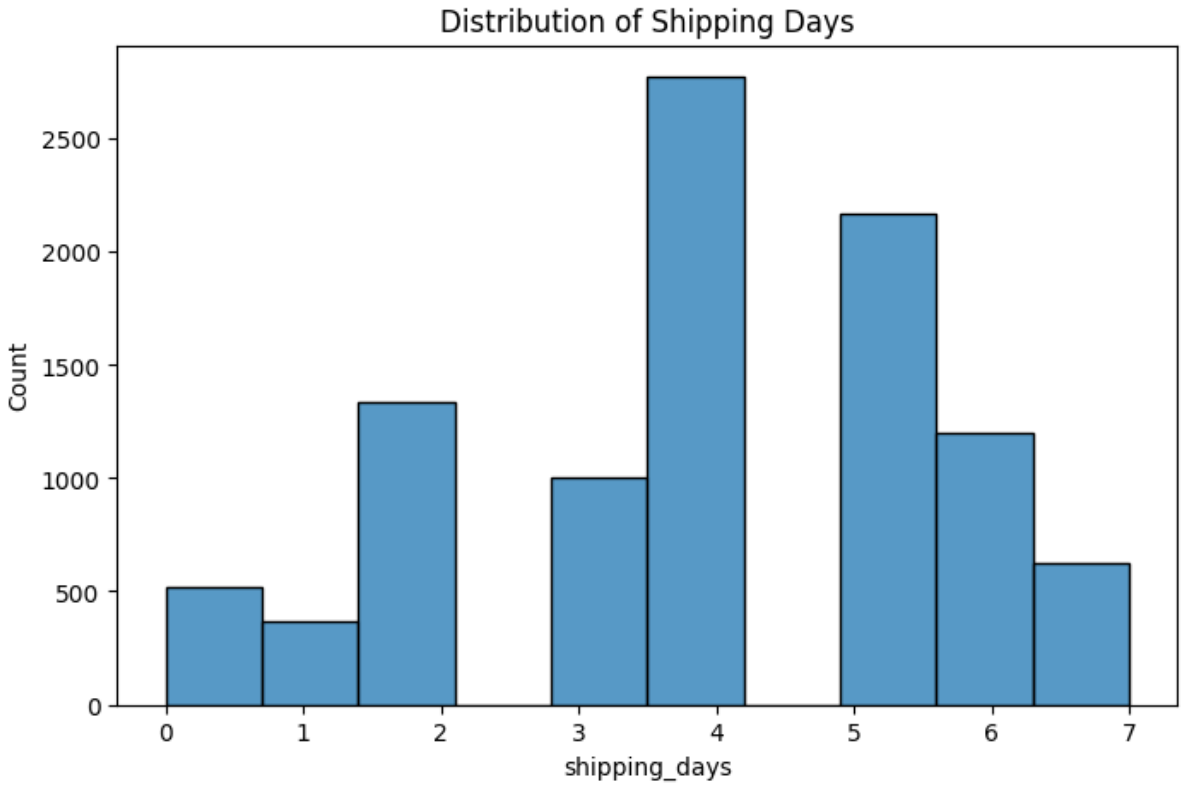
- Sales & Profit by Segment



- Scatterplot: Profit vs Discount



- Histogram: Shipping Days Distribution



Insights

West region has the highest sales & profit.
Consumer segment contributes the most to profit.
Technology category performs best; Furniture struggles.
Higher discounts (>30%) reduce profit.
Average shipping time ~3.96 days — acceptable.

Recommendations

Regional Strategies

Continue to **invest in West & East regions**, which already perform strongly.

Focus on improving sales & profitability in the South region, which currently lags behind.

Investigate what works well in the West/East and replicate best practices in the South.

Customer Segments

Maintain strong engagement with the **Consumer segment**, as it's the most profitable.

Design targeted promotions & outreach for the **Home Office segment**, which is underperforming.

Product Categories

Reassess the **Furniture category**, which shows high sales but low/negative profits.

- Consider optimizing costs, renegotiating supplier contracts, or adjusting prices.
  Promote high-margin **Technology products** more aggressively.
  Reduce focus (or redesign pricing) for loss-making **sub-categories** like Tables & Bookcases.

Discount Policy

Discounts above ~30% lead to significant profit erosion — cap discounts at **20– 30% max**.
Introduce smarter, targeted discounting instead of blanket discounts.

Logistics

Maintain current shipping efficiency (~3–4 days on average), which aligns with customer expectations.
Continue monitoring shipping times & look for incremental improvements.

Tools & Skills Applied

- **Data Cleaning & Preprocessing:** pandas
- **Analysis & Aggregation:** pandas
- **Visualization:** matplotlib, seaborn
- **Statistical Insights:** correlation analysis

Skills: Data cleaning, EDA, descriptive statistics, business insights, storytelling with data.