adv-experiment-1

August 13, 2024

Name: Saad Surve
UID: 2021300131
Class: BE Comps

1 Dataset

You can view the dataset from this link.

2 Description

The dataset contains sales records from a Superstore, detailing orders, products, customers, and financial metrics like sales and profit. It provides insights into which products, regions, categories, and customer segments are most and least profitable.

3 Metadata

- **Row ID**: Unique ID for each row.
- **Order ID**: Unique Order ID for each customer.
- **Order Date**: Date when the order was placed.
- **Ship Date**: Date when the product was shipped.
- **Ship Mode**: Shipping mode chosen by the customer.
- **Customer ID**: Unique ID to identify each customer.
- **Customer Name**: Name of the customer.
- **Segment**: The segment to which the customer belongs.
- **Country**: Country of residence of the customer.
- **City**: City of residence of the customer.
- **State**: State of residence of the customer.
- **Postal Code**: Postal code of the customer's address.
- **Region**: Region where the customer belongs.
- **Product ID**: Unique ID of the product.
- **Category**: Category of the product ordered.
- **Sub-Category**: Sub-category of the product ordered.
- **Product Name**: Name of the product.
- **Sales**: Sales amount of the product.
- **Quantity**: Quantity of the product ordered.
- **Discount**: Discount provided on the product.
- **Profit**: Profit or loss incurred from the sale of the product.

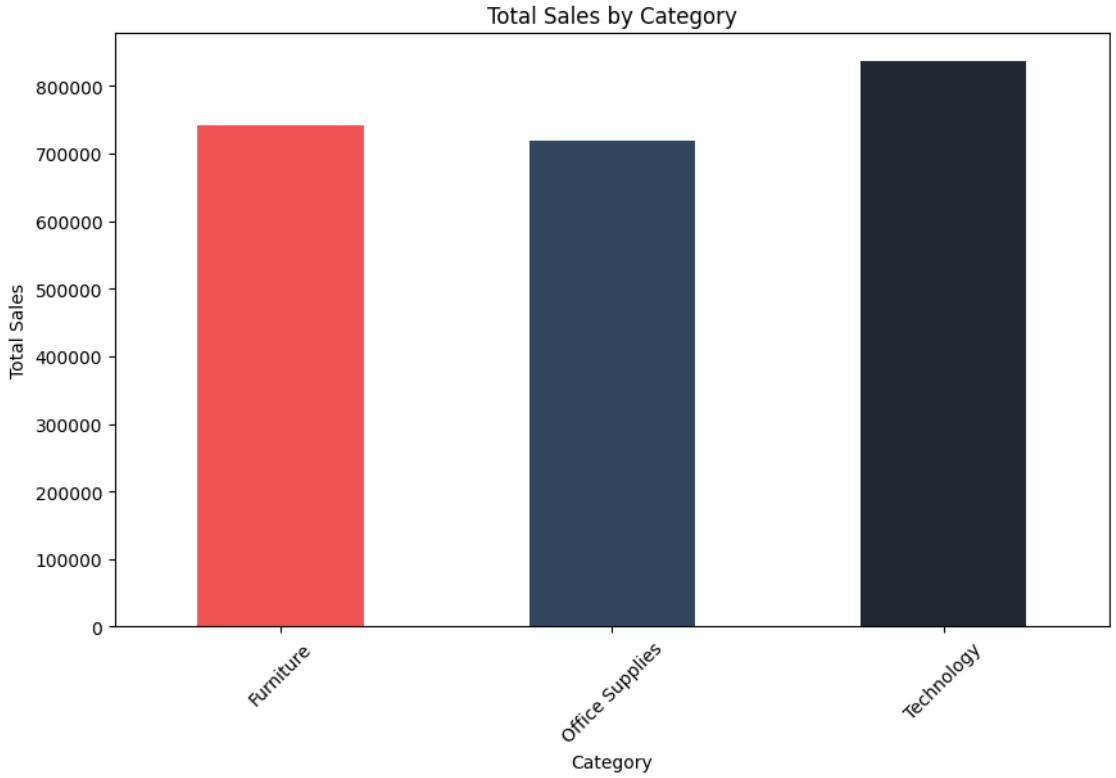
[27]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Read the data
data = pd.read_csv('Sample - Superstore.csv',encoding='windows-1252')
```

4 BAR CHART

[52]:
```
category_sales = data.groupby('Category')['Sales'].sum()

plt.figure(figsize=(10, 6))
category_sales.plot(kind='bar', color=['#F05454','#30475E','#222831'])
plt.title('Total Sales by Category')
plt.xlabel('Category')
plt.ylabel('Total Sales')
plt.xticks(rotation=45)
plt.show()
```



4.1 Observation

The bar chart shows the total sales by category. Among the three categories:

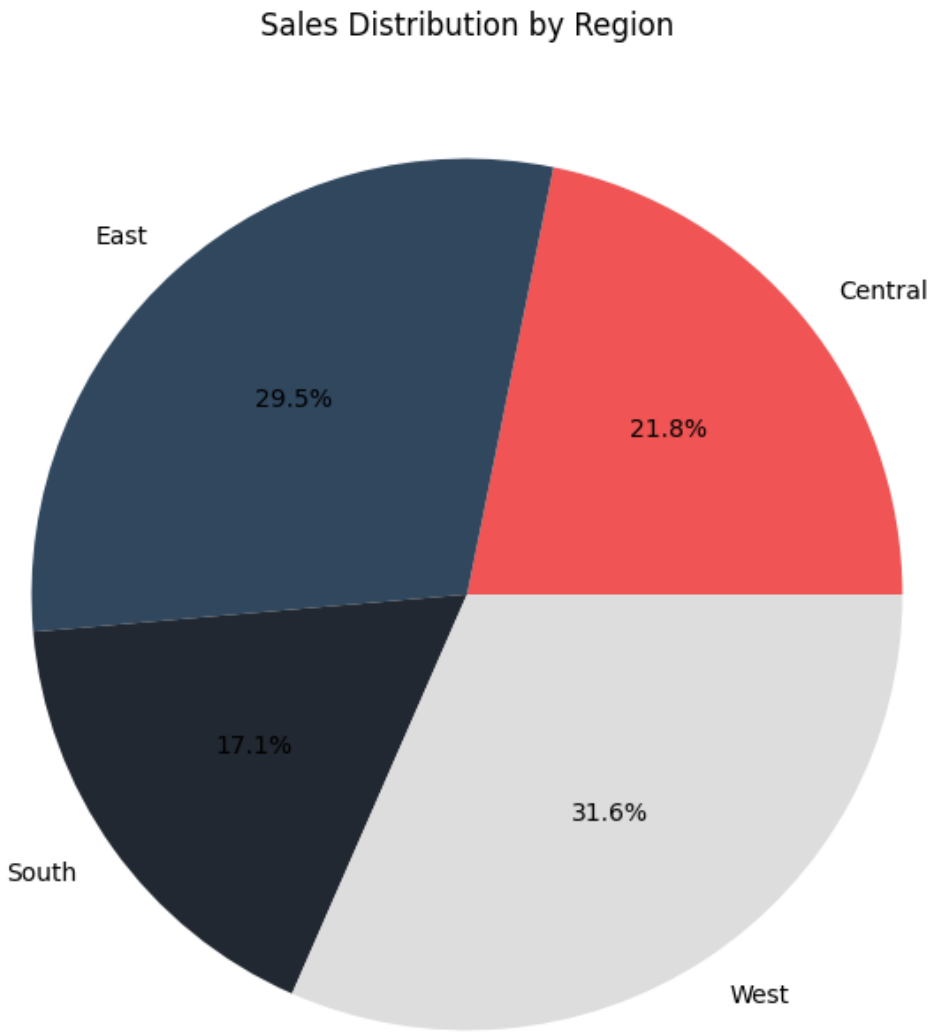
- **Technology** has the highest total sales, slightly surpassing the other categories.
- **Office Supplies** and **Furniture** have similar sales figures, with Office Supplies slightly lagging behind Technology.
- **Furniture** ranks third in terms of total sales but is still quite close to the other two categories.

This suggests that while Technology is the leading category in sales, all three categories contribute significantly to overall sales.

5 PIE CHART

[54]:
```
region_sales = data.groupby('Region')['Sales'].sum()

# Creating the pie chart
plt.figure(figsize=(8, 8))
region_sales.plot(kind='pie', autopct='%1.1f%%',␣
 ↪colors=['#F05454','#30475E','#222831','#DDDDDD'])
plt.title('Sales Distribution by Region')
plt.ylabel('')
plt.show()
```



5.1 Observation

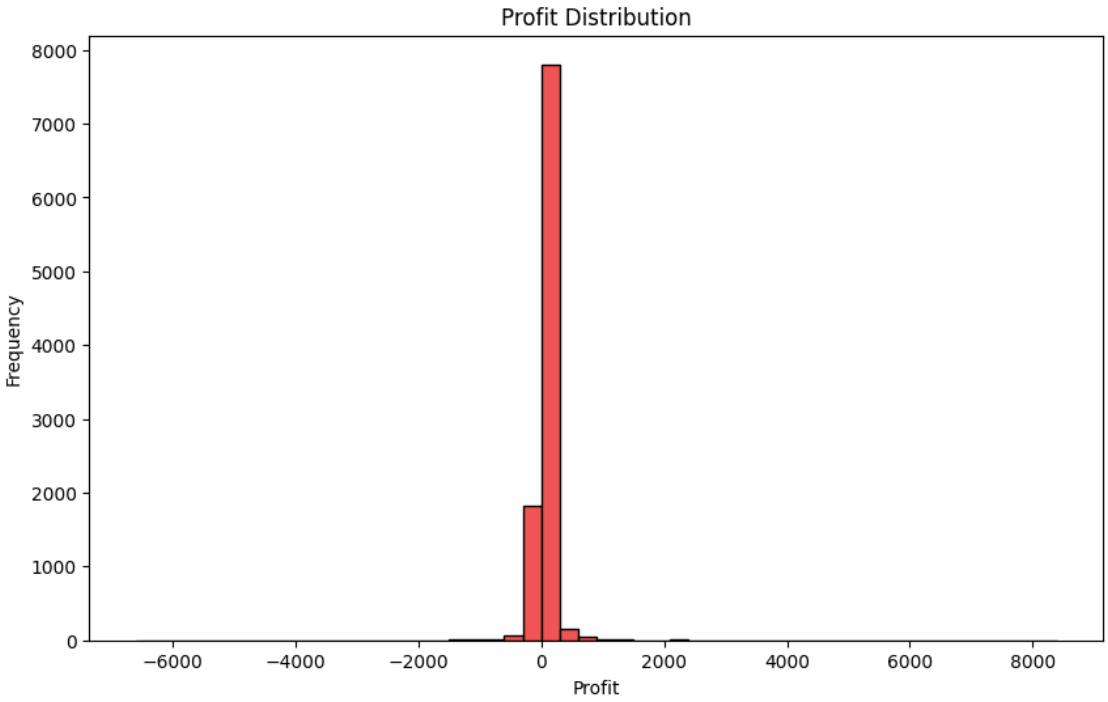
The pie chart illustrates the sales distribution across different regions:

- **West** holds the largest share of total sales at **31.6%**.
- **East** follows closely with **29.5%** of the total sales.
- **Central** accounts for **21.8%**, making it the third-largest contributor.
- **South** has the smallest share, contributing **17.1%** to the total sales.

This suggests that the West and East regions are the most significant contributors to overall sales, while the South region contributes the least.

6 HISTOGRAM

[56]:
```
plt.figure(figsize=(10, 6))
plt.hist(data['Profit'], bins=50, color='#F05454', edgecolor='black')
plt.title('Profit Distribution')
plt.xlabel('Profit')
plt.ylabel('Frequency')
plt.show()
```



6.1 Observation

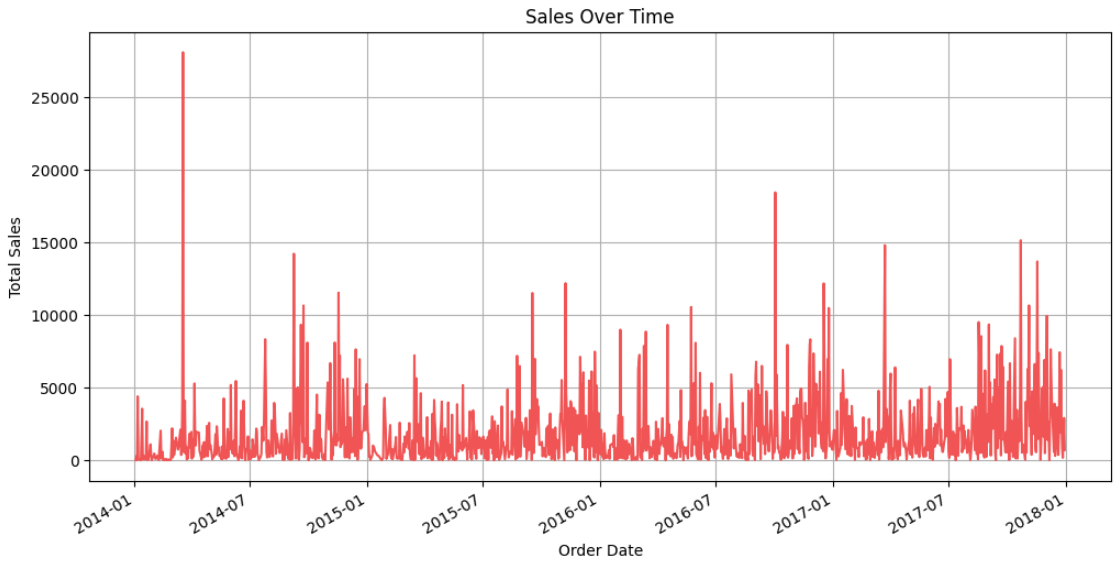
The graph shows a distribution of profits that is highly concentrated around zero, with the majority of the data points falling within a narrow range. There are very few extreme profit values, both positive and negative, as indicated by the small number of occurrences far from the center. This suggests that most of the observations have minimal profit, with rare instances of significant profit or loss.

7 TIMELINE CHART

[57]:
```
# Converting 'Order Date' to datetime format
data['Order Date'] = pd.to_datetime(data['Order Date'])

# Grouping by Order Date and summing Sales
sales_over_time = data.groupby('Order Date')['Sales'].sum()
```

```
# Creating the timeline chart
plt.figure(figsize=(12, 6))
sales_over_time.plot(kind='line', color='#F05454')
plt.title('Sales Over Time')
plt.xlabel('Order Date')
plt.ylabel('Total Sales')
plt.grid(True)
plt.show()
```



7.1 Observation:

The graph shows total sales over time from 2014 to 2018. Sales exhibit significant variability, with frequent spikes and drops. Notably, there are some extreme peaks, particularly around mid-2014 and early 2016, where total sales reach above 20,000 units. Overall, the data suggests fluctuating sales trends without a clear long-term upward or downward pattern, though certain periods experienced higher sales activity.

8 SCATTER PLOT

[55]:
```
# Creating the scatter plot for Sales vs. Profit
plt.figure(figsize=(10, 6))
plt.scatter(data['Sales'], data['Profit'], alpha=0.5, color='#F05454')
plt.title('Sales vs. Profit')
plt.xlabel('Sales')
plt.ylabel('Profit')
plt.grid(True)
plt.show()
```

8.1 Observation:

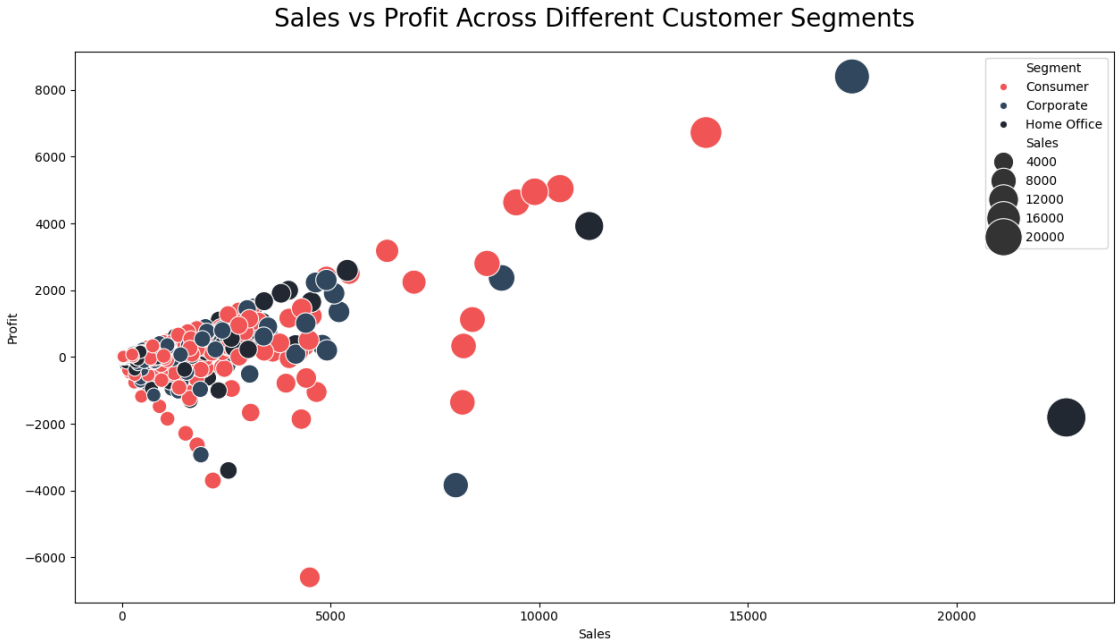
The scatter plot shows a general positive correlation between sales and profit, with most data points clustered in the lower sales range (0-5000). However, there are notable outliers, including some instances of high sales with negative profits, and a few cases of very high profits at moderate to high sales levels.

9 BUBBLE PLOT

[49]:
```
df_scatter = data[['Sales','Profit','Segment']]

plt.figure(figsize=[15,8])
```

```
sns.scatterplot(x=df_scatter['Sales'], y=df_scatter['Profit'],␣
 ↪hue=df_scatter['Segment'], palette=['#F05454','#30475E','#222831'],␣
 ↪size=df_scatter["Sales"], sizes=(100,1000), legend='auto')
plt.title("Sales vs Profit Across Different Customer Segments", size=20, pad=20)
plt.show()
```



9.1 Observation:

This bubble plot illustrates sales vs. profit across different customer segments:

1. The plot uses bubbles instead of points, with bubble size representing sales volume.
2. Three customer segments are color-coded: Consumer (red), Corporate (dark blue), and Home Office (black).
3. Larger bubbles tend to appear in the higher sales and profit ranges, particularly for Corporate and Consumer segments.
4. The largest bubble, representing the highest sales volume, is a Corporate segment transaction with very high profit.
5. There are some large bubbles in negative profit areas, indicating high-volume sales that resulted in losses.