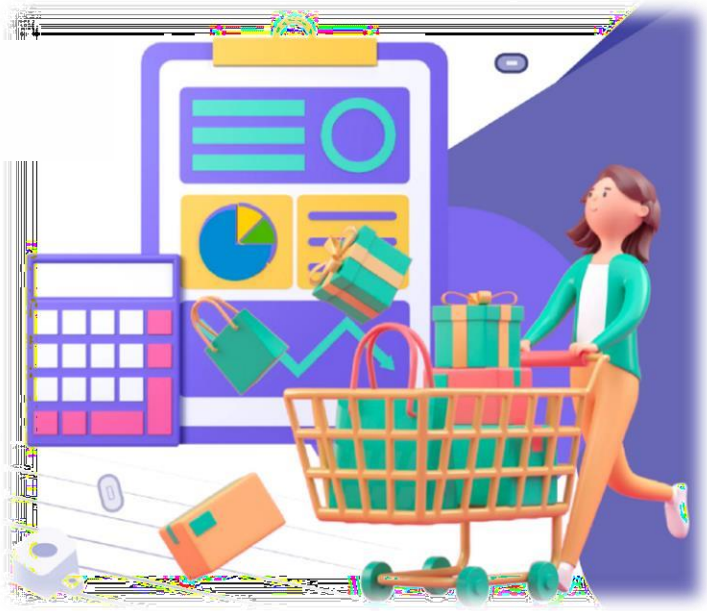

SUPERSTORE SALES ANALYSIS

DEPI Final Project

SUPERVISOR:
Dr. Soha Nagy

Nada Mohamed Abbas
Marwa Omar Muhammed
Marwa Hussein Ahamed Ali
Dalia Mahmoud El Sayed
Mona Ahamed Abdelmonem
Roqia Mohammed Amin

Track: Data Analyst Specialist
Group: YAT200

Table of Contents

▪ Introduction:

Understanding regional variations in consumer behavior is critical for businesses seeking to optimize their product offerings and marketing strategies. Different geographic areas exhibit distinct purchasing patterns, with urban consumers often gravitating toward the latest technology, while rural consumers may prioritize more practical goods, such as tools or farming equipment. As the global marketplace becomes increasingly interconnected, it is imperative for companies to not only identify their target customers but also to recognize the influence of location on purchasing decisions.

This report seeks to analyze sales data to compare consumer preferences between urban and rural areas and between east and west regions. By examining these trends, the report aims to provide actionable insights that can guide businesses in tailoring their strategies to better meet the needs of diverse regional markets.

The project begins with a raw dataset that presents initial challenges due to its ununified and incomplete nature. However, within this data lies valuable information, and through careful analysis and transformation, we aim to uncover insights that shed light on the factors influencing consumer behavior in different geographic settings.

▪ Tools and Technologies:

This project was developed using Python, executed within the Visual Studio environment. Several key libraries were utilized to support data manipulation, visualization, and preprocessing tasks:

- **Pandas**: The Pandas library was employed for efficient data manipulation and analysis. Its flexible data structures, such as Data Frames, allowed for easy handling and transformation of the dataset, including operations such as filtering, grouping, and aggregation.
- **Matplotlib and Seaborn**: For data visualization, we utilized both Matplotlib and Seaborn libraries. Matplotlib provided a foundation for creating basic plots and graphs, while Seaborn enhanced these visualizations with more advanced features, offering aesthetic, informative charts for deeper analysis of trends and relationships within the data.
- **Scikit-learn (Simple Imputer)**: To address missing data, we used the Simple Imputer class from the Scikit-learn library. This tool allowed for efficient handling of incomplete datasets by imputing missing values based on strategies such as mean, median, or mode replacement, ensuring that the dataset remained complete and consistent for analysis.

▪ Data:

Data Description: The dataset used for this analysis comprises sales transaction records from a large retail superstore operating in the United States, covering the period from 2015 to 2018. It captures various aspects of customer orders, including product details, shipping information, and customer demographics. This dataset provides a rich source of information to analyze and compare consumer preferences across different regions of the United States. The dataset was sourced from **Data Analyst Specialist - Project Ideas** and includes multiple columns that track essential sales-related details.

Key Features: The dataset contains the following key columns, each contributing critical insights to the analysis:

- **Row ID**: A unique identifier for each record (Categorical). It is used primarily for indexing and does not play a direct role in the analysis.
- **Order ID**: A unique identifier for each transaction (Categorical). This field helps track individual orders, allowing us to analyze sales at a granular level.
- **Order Date**: The date on which the order was placed (Datetime). This feature enables time-series analysis, helping to identify trends such as seasonality and changes in consumer behavior over time, from 2015 to 2018.
- **Ship Date**: The date the order was shipped (Datetime). This field is useful for analyzing shipping speed and its potential impact on customer satisfaction.
- **Ship Mode**: The shipping method used for the order (Categorical), such as standard class, second class, or same-day delivery. This feature can provide insights into how shipping preferences vary by region or customer segment.
- **Customer ID**: A unique identifier for each customer (Categorical). This feature helps analyze repeat purchasing patterns and customer loyalty across different regions.
- **Customer Name**: The name of the customer (Categorical). While not directly contributing to the analysis, this column can be useful for grouping transactions by individual customers.
- **Segment**: The segment to which the customer belongs (Categorical), such as "Consumer," "Corporate," or "Home Office." This feature is essential for analyzing how purchasing preferences differ across different types of customers.
- **Country**: The country in which the transaction occurred (Categorical). In this dataset, all transactions are from the **United States**. While this feature is constant, it reinforces the geographical focus of the study.
- **City**: The city where the customer is located (Categorical). This feature plays a central role in comparing consumer behavior between urban and rural areas, allowing for location-based segmentation of sales data.

- **Postal Code**: The postal code of the customer's location (Numerical). This feature can be used to group transactions by region and understand geographic patterns at a more granular level.
- **Region**: The region of the United States in which the transaction took place (Categorical), such as "East," "West," "South," and "Central." This feature allows for regional analysis of sales trends and consumer behavior across different parts of the country.
- **Product ID**: A unique identifier for each product (Categorical). This field helps track the sales of specific products and can be useful for inventory management or product popularity analysis.
- **Category**: The main category of the product sold (Categorical), such as "Furniture," "Office Supplies," or "Technology." This feature is crucial for analyzing sales performance across different product categories.
- **Sub-Category**: A more detailed classification within the main product category (Categorical), such as "Chairs," "Binders," or "Phones." This feature allows for a more nuanced analysis of product preferences within each category.
- **Product Name**: The name of the product sold (Categorical). This feature enables product-level sales analysis and helps in identifying the most and least popular items.
- **Sales**: The sales revenue generated from each transaction (Numerical). This feature is essential for assessing the financial performance of different products, regions, and customer segments, and is a primary metric for the analysis.

These features together provide a comprehensive view of the sales process, from the moment the order is placed to the delivery of the product. By analyzing this data, we can gain insights into how consumer preferences vary by region, customer segment, and product type, and how these factors influence overall sales performance.

Data Challenges: During the data preprocessing stage, several challenges were encountered. One of the primary issues was identifying which columns contained missing values, as some fields had incomplete data. Specifically, the **Postal Code** column exhibited missing values, which required further attention. Additionally, duplicate records were present within the dataset, posing the risk of skewing the analysis if not properly addressed. Another challenge involved determining whether certain columns, such as **Region**, represented broad categories or more granular data, necessitating an examination of the unique values within each column. Furthermore, the **Ship Date** and **Order Date** columns displayed inconsistent formatting, complicating temporal analysis. These challenges underscored the need for comprehensive data cleaning and transformation to ensure the accuracy and reliability of the analysis.

▪ Data cleaning:

To prepare the dataset for analysis, a series of data cleaning steps were implemented. The dataset was initially loaded, with the index set to the Row ID for easier manipulation. A thorough examination identified columns with missing values, leading to calculations of the percentage of missing data for each feature. Notably, the Postal Code column contained incomplete entries, specifically in rows corresponding to Vermont. To address these null values, we utilized the Simple Imputer from Scikit-learn with the "constant" strategy, filling them with the value 05401. Since Postal Codes begin with "0" and do not require numerical operations, the values were cast to strings to ensure proper formatting. The dataset was further evaluated for unique values to differentiate between categorical and granular data. Duplicate entries were identified and removed to maintain data integrity. Additionally, the Order Date and Ship Date columns were standardized due to inconsistent formats, converting them to a unified datetime format. The cleaned dataset was then saved for further analysis, ensuring accuracy and reliability. For a detailed overview of the data cleaning process, the corresponding code can be found in the appendix.

▪ Visualization and Driven-Insights:

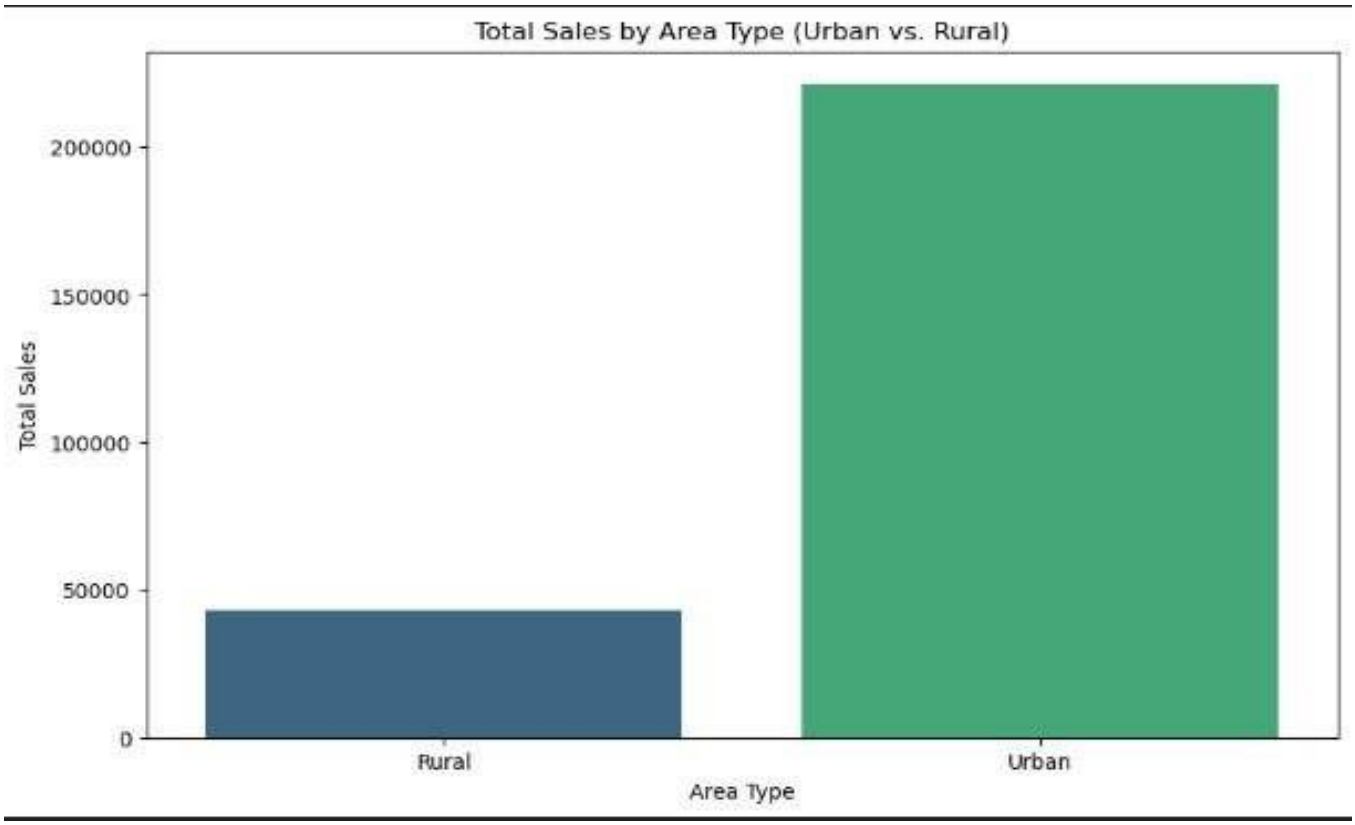

Chart (1)

Based on the chart (1), the total sales disparity shows that urban areas significantly outperform rural areas in terms of total sales. This means urban areas contribute the vast majority of the overall sales for the store, indicating a heavy reliance on urban sales. To diversify the sales base, efforts could be made to increase sales in rural areas, potentially through targeted promotions, improved logistics, or localized product offerings.

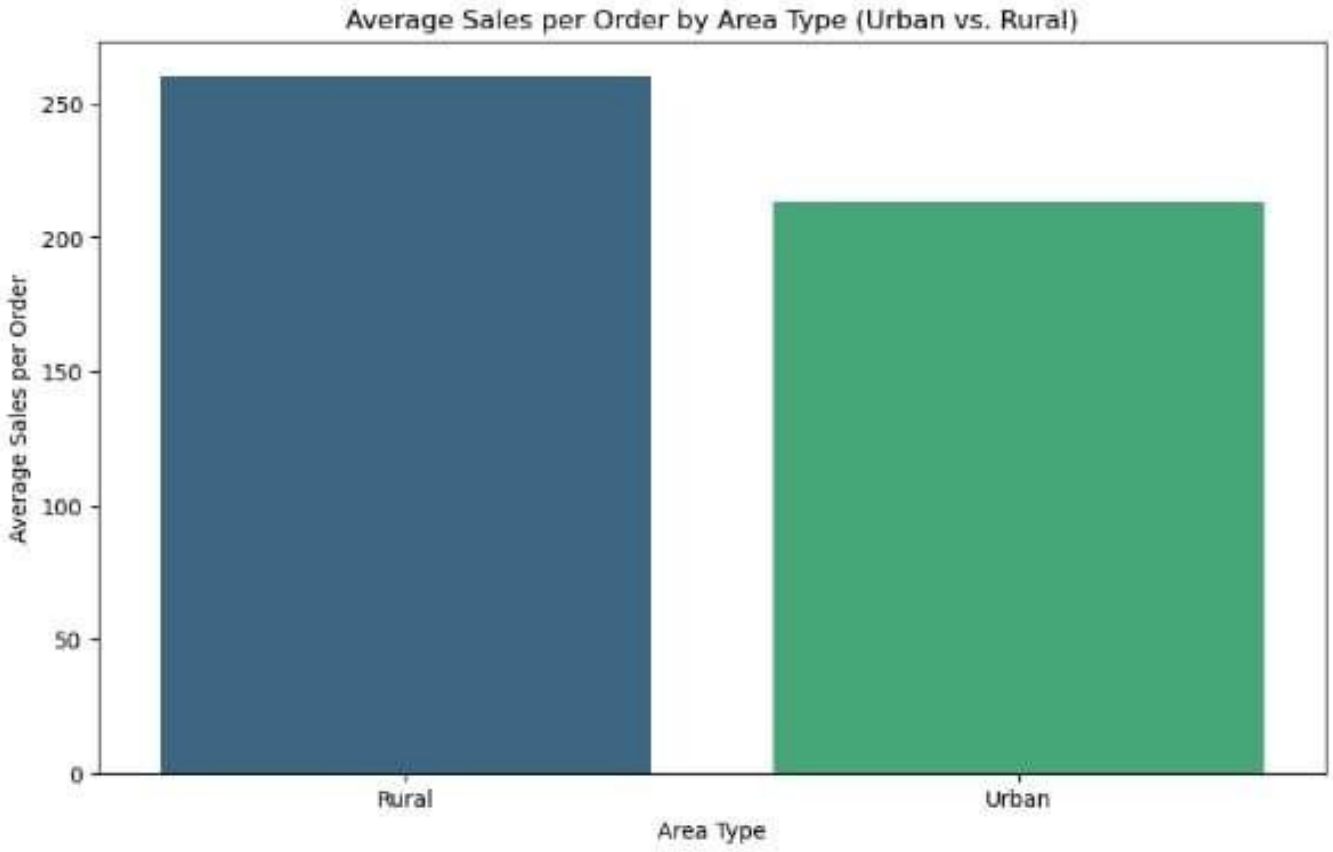

Chart (2)

Based on chart (2), Rural areas have a higher average sales per order. In contrast, urban areas have a lower average sales per order. Despite the lower total sales in rural areas, the higher average sales per order present an interesting growth opportunity. The store could focus on enhancing this by possibly expanding product offerings tailored to rural customers.

Urban areas show a steady stream of smaller orders, possibly driven by accessibility and purchasing convenience. Increasing the average order size in urban regions through bundled promotions or incentives could further boost revenue.

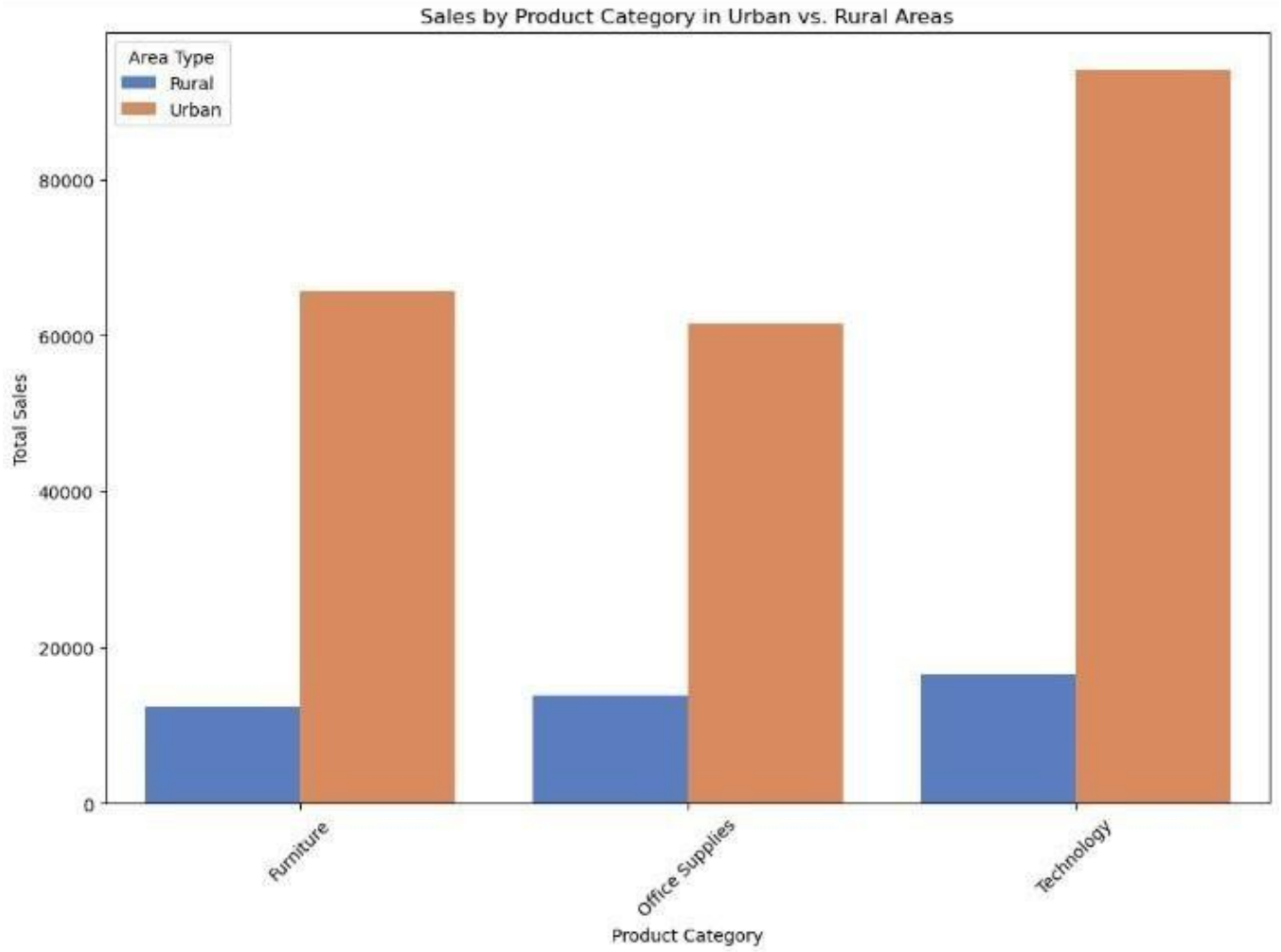

Chart (3)

Based on chart (3), urban areas exhibit higher sales across all product categories compared to rural areas. Technology products have the largest sales gap between urban and rural regions. While urban areas also lead in furniture and office supplies sales, the differences are less significant. Businesses can prioritize urban areas for maximum sales but should also consider strategies to tap into the rural market potential. Product differentiation, tailored marketing, and appropriate distribution channels are essential for success in both regions.

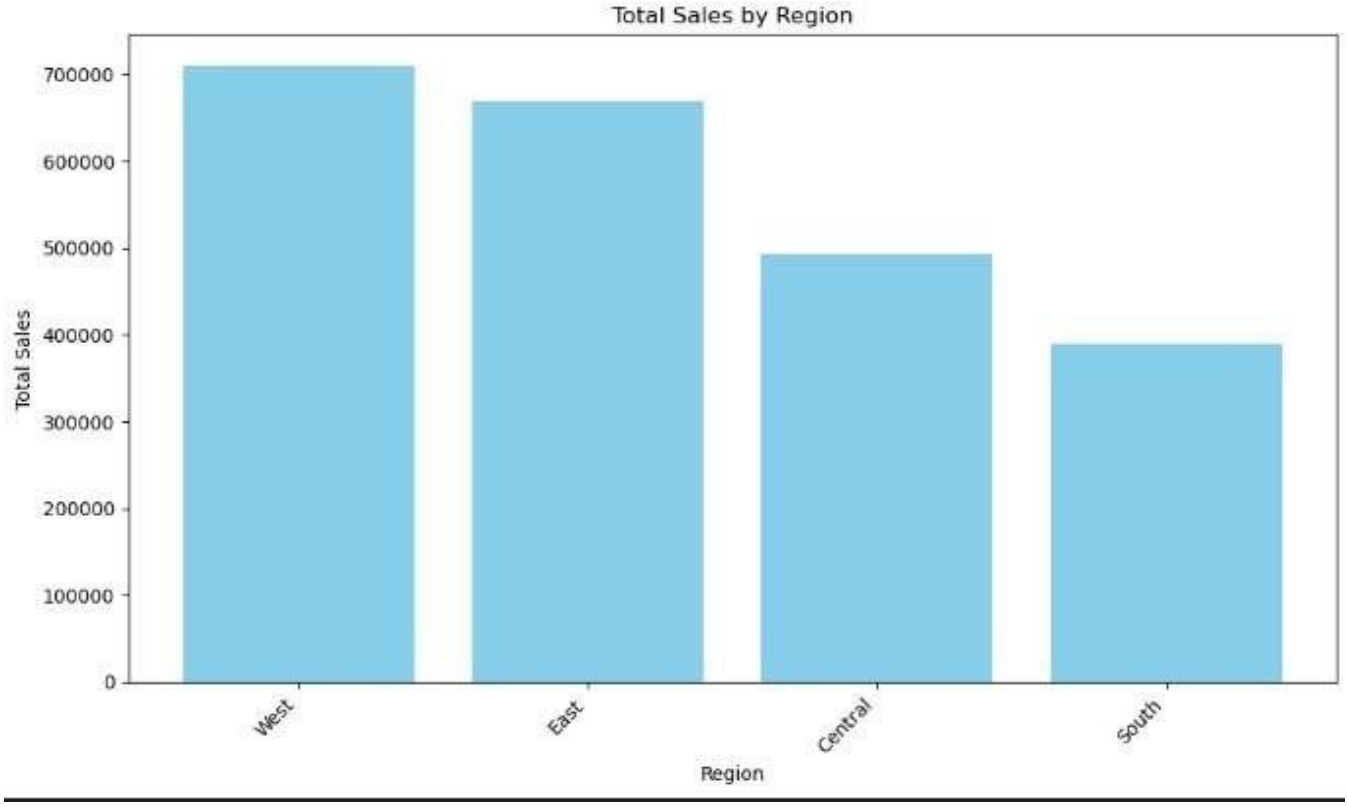

Chart (4)

Based on the chart (4), here are some insights: The West region has the highest total sales, exceeding 700,000, making it the top-performing region in terms of sales. The East region follows closely, with sales slightly under 700,000, still demonstrating strong performance. The Central region shows a notable drop in sales, totaling around 500,000, placing it as the third-best region. Lastly, the South region has the lowest sales, falling below 500,000, indicating potential challenges. We need to focus on the South region, exploring the reasons behind its performance and developing strategies to boost sales, which could be beneficial. Additionally, it is essential to sustain performance in the West and East regions, as these areas are performing well; continued engagement offers, or targeted marketing could ensure that they maintain their high sales figures. Furthermore, we must investigate the Central region, as there is a significant gap between it and the two top regions, which could be addressed by tailoring sales strategies or promotions to encourage growth.

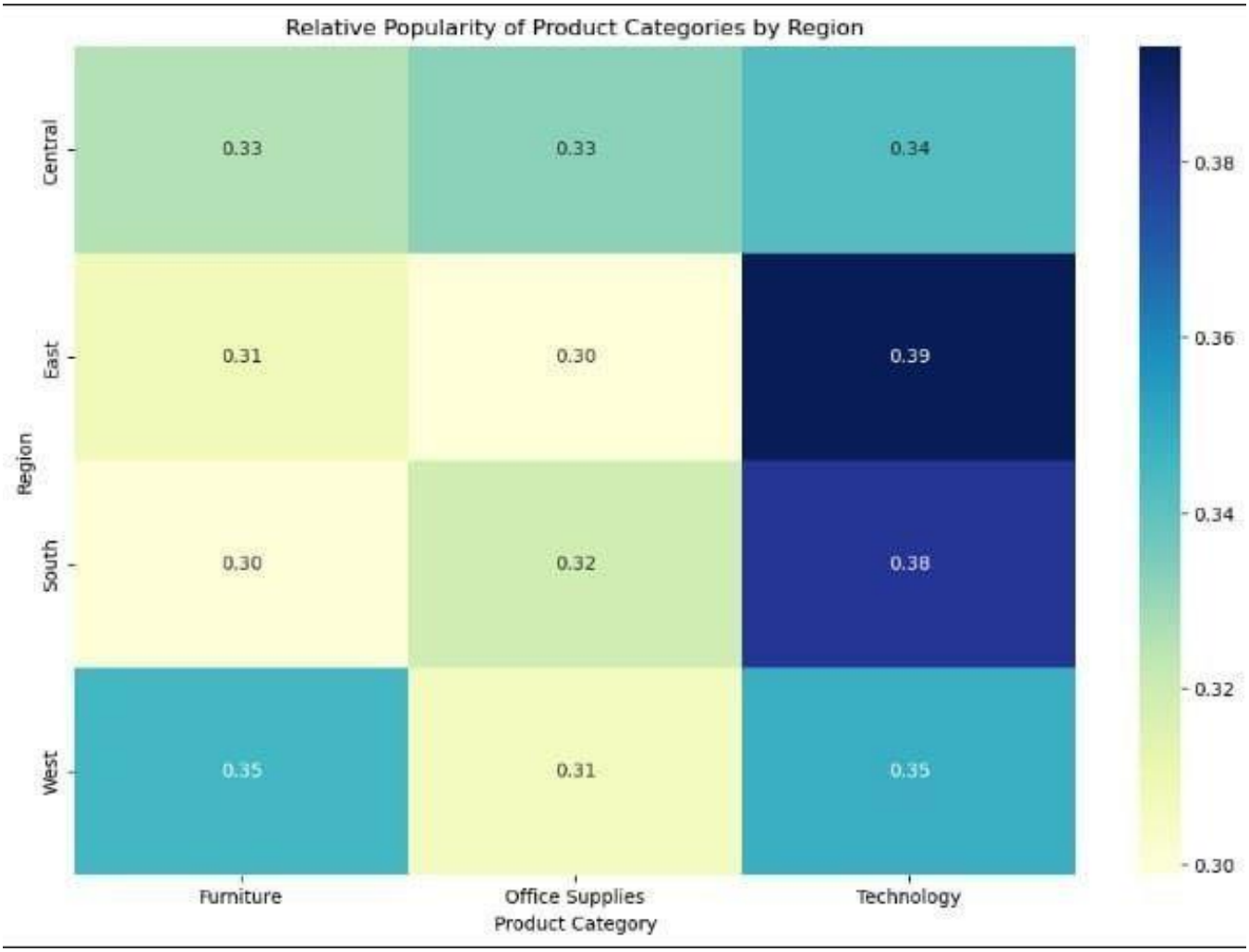

Chart (5)

The heatmap shows that the East region exhibits a strong preference for technology, with a lower interest in office supplies. In contrast, the West region demonstrates a more balanced interest across all categories, with a slight preference for furniture and technology. The Central region shows relatively balanced demand across the categories, with no strong spikes. Lastly, the South region indicates a higher interest in technology but comparatively lower interest in furniture and office supplies.

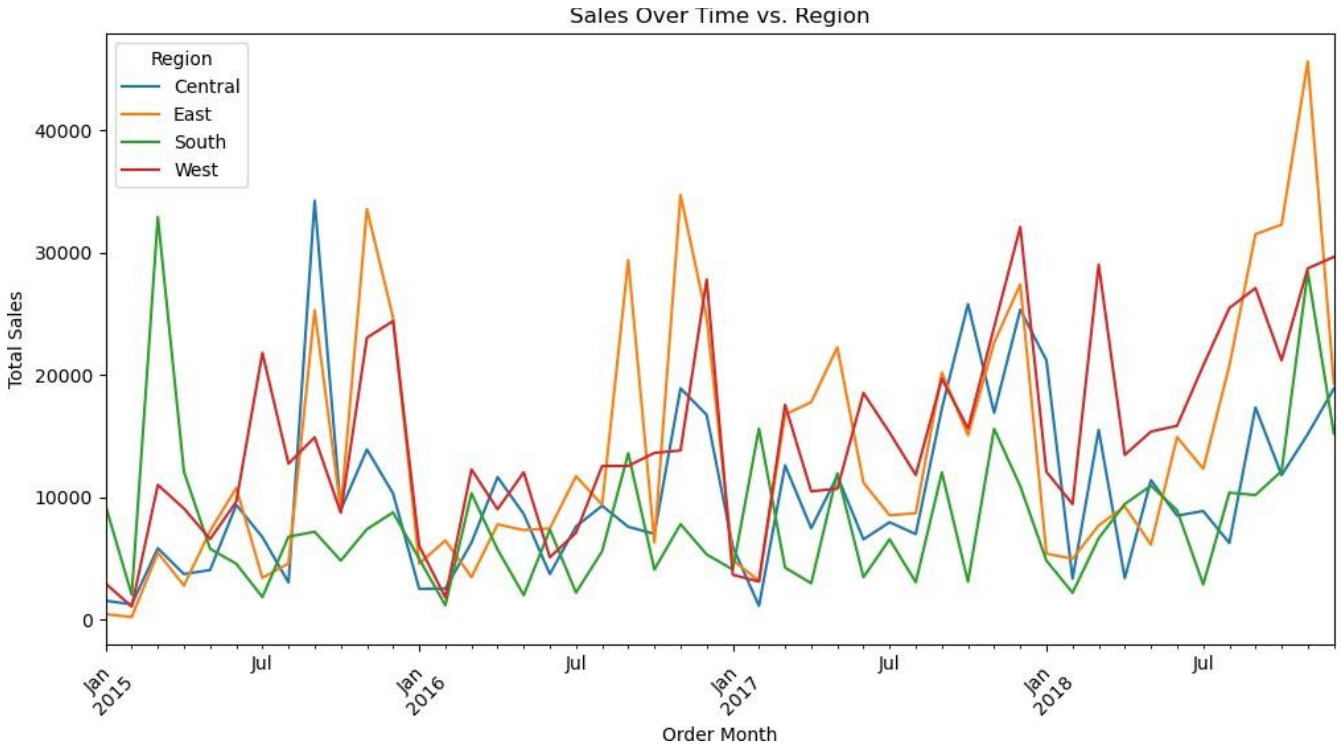

Chart (6)

Based on the chart (6), here are some insights: In the East region, sales have shown fluctuations but demonstrate an overall upward trend, with a possible hint of seasonality. The highest point in sales was observed in November 2018. The South region, on the other hand, displays highly volatile sales with significant peaks and troughs. While there is a slight upward trend, the fluctuations are much more pronounced than in the East, with the peak occurring in March 2015.

Similarly, the Central region exhibits considerable volatility, with sales characterized by peaks and troughs. Despite this, an overall upward trend is visible, although the volatility remains high. The peak for this region occurred in September 2015. In the West region, sales have experienced significant fluctuations but maintain an upward trajectory, with a slightly more pronounced seasonal pattern compared to the East. the West region reached its sales peak in December 2017.

When comparing the regions, the South and Central regions show the highest levels of volatility, with dramatic swings in sales. In contrast, the East and West regions exhibit less extreme fluctuations, suggesting more stability. Seasonal patterns, while present in all regions, appear more pronounced in the East and West. Overall, all regions display a positive upward trend in sales, indicating growth over the analyzed period. However, the rate of growth and the stability of sales trends differ, with the East and West regions showing more consistent growth, while the South and Central regions experience more volatility.

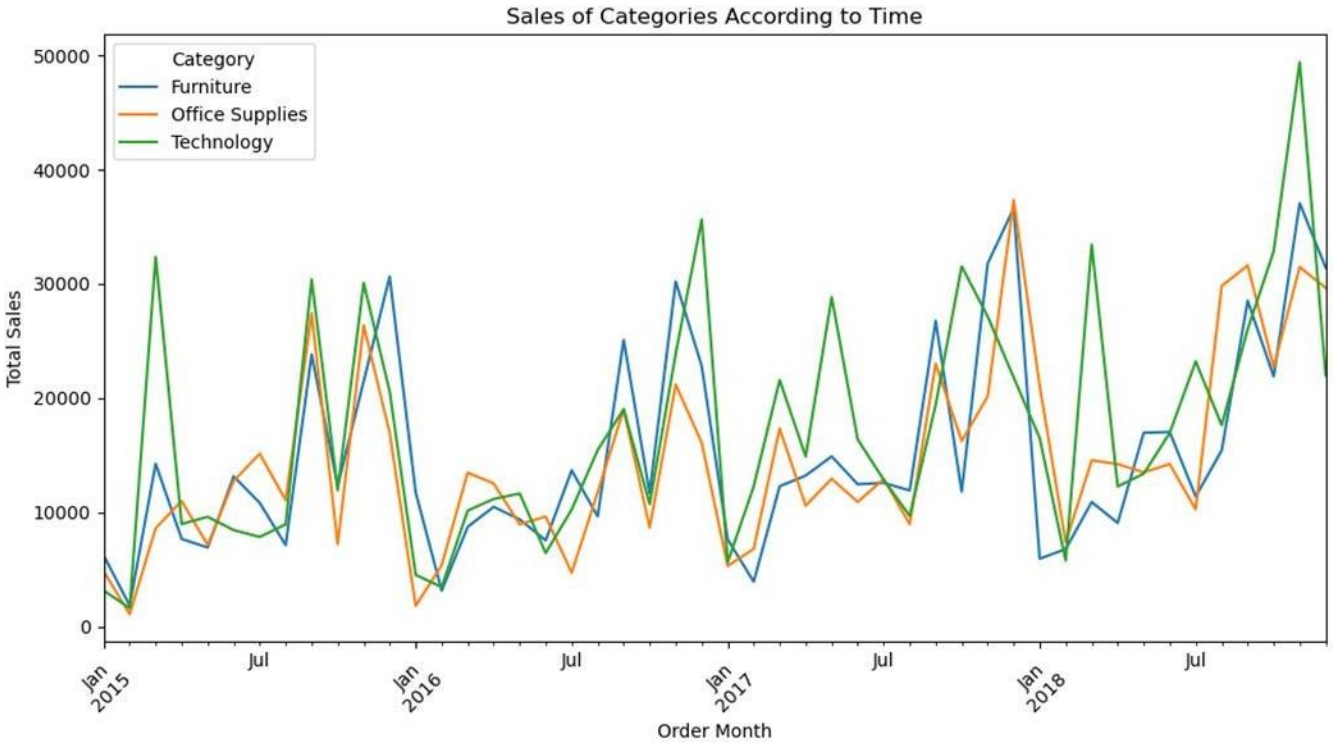

Chart (7)

Chart (7) illustrates sales trends for Furniture, Office Supplies, and Technology from 2015 to 2018. Each category shows fluctuations, with peaks around the year-end, suggesting seasonality in consumer demand. Furniture sales are more stable but spike occasionally, likely due to bulk purchases by businesses. Office Supplies follow a consistent upward trend, reflecting steady demand. Technology, however, shows sharp increases, particularly toward 2018, indicating growing investment in tech products. These insights suggest that the company should prepare for seasonal sales surges, especially in technology, and align inventory and marketing strategies accordingly.

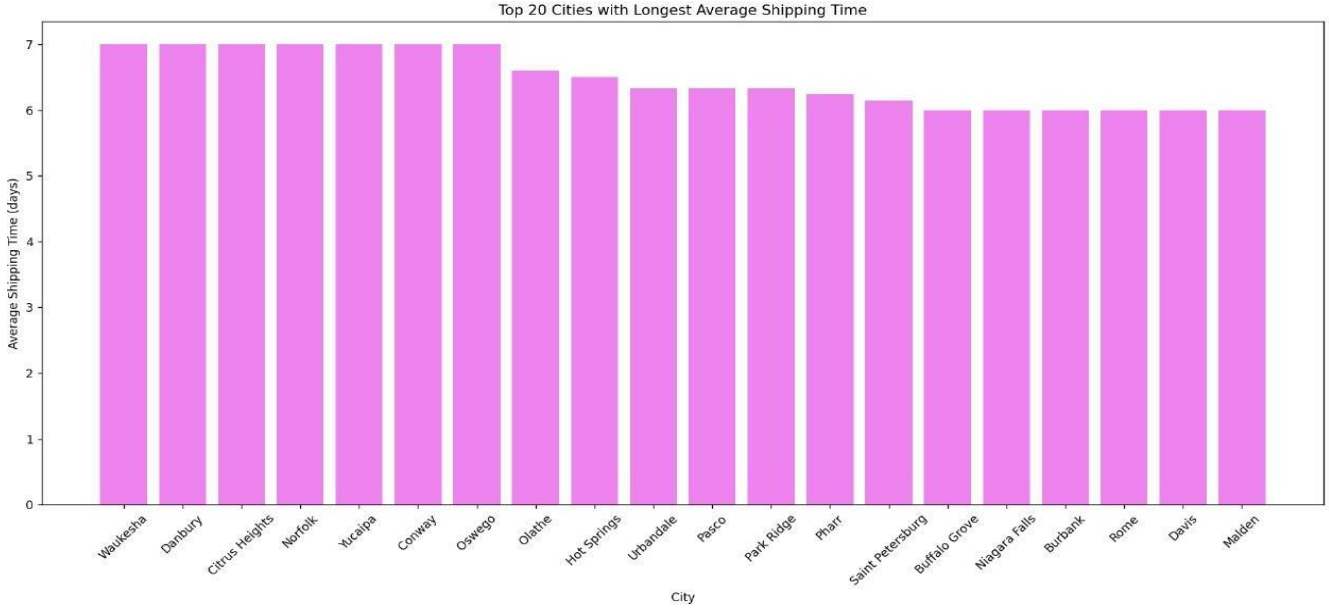

Chart (8)

chart (8) shows the top 20 cities with the longest average shipping times, with values ranging from 6 to 7 days. Most of these cities appear to be rural or suburban, indicating that less urbanized areas face longer shipping delays, likely due to limited infrastructure and distance from major distribution hubs. Urban cities, while present, show slightly faster shipping times. Improving logistics in rural areas could reduce these delays.

▪ Recommendations:

To optimize the business's performance across different geographic areas and consumer segments, the following strategies can enhance sales and address market disparities between urban and rural areas as well as regional differences:

Maximize Sales in Urban Areas:

Urban areas account for the majority of total sales, particularly in high-demand categories like technology. Given the sales spikes toward the year-end and the increasing trend in technology, the business should enhance its product offerings in these regions, focusing on innovative tech products. Targeted marketing campaigns, especially during peak seasons, can drive average order sizes. Introducing bundled promotions, loyalty programs, or special offers can encourage larger purchases, further increasing revenue.

Unlock the Potential in Rural Markets:

Although rural areas contribute less to total sales, they exhibit higher average sales per order. To capitalize on this, the business should expand product offerings tailored to rural consumers, focusing on practical or high-value items. Given the longer shipping times observed in these areas, strengthening distribution channels to improve logistics will be essential. Developing marketing campaigns that highlight rural-specific promotions can cater to the distinct preferences of rural customers and stimulate purchasing frequency.

Address Regional Sales Disparities:

The West region leads in sales performance, while the South lags behind. To improve performance in the South, the business should investigate the reasons behind its underperformance and implement targeted promotions. Meanwhile, sustaining the strong performance in the West and East regions through continued customer engagement and seasonal promotions is crucial. In the Central region, characterized by moderate but volatile sales, consistent marketing efforts and strategic product offerings can stabilize demand and drive growth.

Reduce Sales Volatility:

Managing sales volatility is critical, especially in regions with significant fluctuations. Introducing consistent promotions during off-peak times, particularly where extreme peaks and troughs are observed, can help smooth sales patterns. Seasonal strategies should be employed to encourage steady year-round growth, especially in technology, which shows sharp increases in demand.

Optimize Product Differentiation and Distribution:

Given that urban consumers drive higher sales across all categories, especially in technology, the business should continue to prioritize tech sales in urban areas while expanding these offerings in rural regions. Tailoring product offerings to meet the unique needs of each region ensures alignment with local consumer demand. Additionally, refining distribution strategies is crucial to address the logistical challenges faced in rural areas, which often experience longer shipping times. Improving logistics in these regions will enhance service levels and boost sales potential.

By following these recommendations, the business can diversify its sales base, drive growth in underperforming regions, and create a sustainable and balanced approach to revenue generation. This strategy will strengthen performance in high-performing regions while unlocking new opportunities in areas with untapped potential, ultimately leading to a more resilient and profitable business model.

▪ Conclusion:

This analysis has revealed critical insights into the consumer preferences and sales patterns across different geographic regions in the United States, emphasizing the contrasts between urban and rural markets as well as regional sales disparities. Urban areas dominate total sales, particularly in technology, while rural regions show higher average sales per order, presenting growth opportunities. Regionally, the West outperforms other areas, while the South struggles, indicating the need for tailored strategies to bridge these gaps. By enhancing product differentiation, optimizing distribution, and addressing sales volatility, businesses can effectively boost performance across diverse markets.

With targeted promotions and strategic product offerings in both urban and rural areas, as well as addressing regional challenges, companies can achieve more balanced, sustainable growth. The combination of regional focus, logistical improvements, and marketing adjustments can unlock significant opportunities in underperforming areas while sustaining growth in stronger markets.

- **Appendix:**

```
import pandas as pd
from sklearn.impute import SimpleImputer
```

```
df= pd.read_csv("Superstore Sales Dataset.csv")
```

```
df.head()
```

```
df.set_index('Row ID', inplace= True)
```

```
df.head()
```

```
df.shape
```

```
total_rows = df.shape[0]

for col in df.columns:
    nan_count = df[col].isnull().sum()
    percentage_nan = (nan_count / total_rows) * 100
    print(f"{col}: {percentage_nan:.2f}%")
```

```
df[df.isnull().any(axis=1)].head()
```

```
df[df['City'] == 'Burlington']
```

```
df['Postal Code'].dtype
```

```
imputer = SimpleImputer(strategy= 'constant', fill_value= 5401)
```

```
df['Postal Code'] = imputer.fit_transform(df[['Postal Code']])
```

```
df['Postal Code'] = df['Postal Code'].astype(int)
```

```
df['Postal Code'] = df['Postal Code'].astype(str)
```

```
df['Postal Code'] = df['Postal Code'].str.replace('5401', '05401')
```

```python
df['Postal Code'].dtype
```

Python

```python
df[df['City'] == 'Burlington']
```

Python

```python
for column in df.columns:
    print(f"Unique values in {column}:")
    print(df[column].unique().shape)
    print()
```

Python

```python
df= df.drop(df[df.duplicated()].index)
```

Python

```python
df['Order Date'] = pd.to_datetime(df['Order Date'], format= "%d/%m/%Y")
df['Ship Date'] = pd.to_datetime(df['Ship Date'], format= "%d/%m/%Y")
```

Python

```python
df.head()
```

Python

```python
df.to_csv("SuperStore Sales_cleaned.csv")
```

Python

```python
# Group by Region and calculate total Sales
sales_by_region = df.groupby('Region')['Sales'].sum().reset_index()

# Sort the regions by Sales in descending order
sales_by_region = sales_by_region.sort_values(by='Sales', ascending=False)

# Plotting the bar chart
plt.figure(figsize=(10, 6))
plt.bar(sales_by_region['Region'], sales_by_region['Sales'], color='skyblue')
plt.xlabel('Region')
plt.ylabel('Total Sales')
plt.title('Total Sales by Region')
plt.xticks(rotation=45, ha='right')  # Rotate region names for better readability
plt.tight_layout()  # Adjust layout to ensure everything fits without overlap
plt.show()
```

```python
# Group by Region and Year-Month, then sum Sales
sales_trend = df.groupby(['Region', 'Year-Month'])['Sales'].sum().reset_index()

# Pivot the table to have Year-Month as columns and Region as rows
sales_trend_pivot = sales_trend.pivot(index='Region', columns='Year-Month', values='Sales')

# Calculate the percentage change month over month
sales_trend_pct_change = sales_trend_pivot.pct_change(axis=1)

# Identify regions with consistently declining sales (e.g., negative growth over multiple periods)
# For simplicity, we'll define "consistently declining" as having negative growth for three or more consecutive months
declining_regions = sales_trend_pct_change.apply(lambda row: (row < 0).sum() >= 3, axis=1)
declining_regions = declining_regions[declining_regions].index.tolist()

print("Regions with consistently declining sales:", declining_regions)

# Visualize the trends for declining regions
for region in declining_regions:
    plt.figure(figsize=(10, 6))
    sales_trend_pivot.loc[region].plot(marker='o')
    plt.title(f'Sales Trend Over Time for {region}')
    plt.xlabel('Time (Year-Month)')
    plt.ylabel('Sales')
    plt.xticks(rotation=45)
    plt.grid(True)
    plt.tight_layout()
    plt.show()
```

```python
# Group by Region and Category, then sum Sales
category_sales_by_region = df.groupby(['Region', 'Category'])['Sales'].sum().reset_index()

# Pivot the table to have Category as columns and Region as rows
category_sales_pivot = category_sales_by_region.pivot(index='Region', columns='Category', values='Sales')

# Normalize the sales data to see relative popularity within each region
category_sales_normalized = category_sales_pivot.div(category_sales_pivot.sum(axis=1), axis=0)

# Plotting heatmap to visualize category popularity by region
plt.figure(figsize=(12, 8))
sns.heatmap(category_sales_normalized, annot=True, cmap="YlGnBu", fmt=".2f")
plt.title('Relative Popularity of Product Categories by Region')
plt.xlabel('Product Category')
plt.ylabel('Region')
plt.show()
```

```python
# Example classification of cities into Urban or Rural
city_classification = {
    'New York': 'Urban',
    'Los Angeles': 'Urban',
    'Chicago': 'Urban',
    'Springfield': 'Rural',
    'Boise': 'Rural',
    # Add all relevant cities here
}

# Create a new column 'Area Type' based on city classification
df['Area Type'] = df['City'].map(city_classification)

# Check for unclassified cities
unclassified_cities = df[df['Area Type'].isna()]['City'].unique()
if len(unclassified_cities) > 0:
    print(f"Unclassified cities: {unclassified_cities}")

# Group by Area Type and calculate total and average sales
sales_by_area_type = df.groupby('Area Type')['Sales'].agg(['sum', 'mean', 'count']).reset_index()

# Rename columns for clarity
sales_by_area_type.columns = ['Area Type', 'Total Sales', 'Average Sales per Order', 'Number of Orders']



# Plot the total sales by Urban vs. Rural
plt.figure(figsize=(10, 6))
sns.barplot(x='Area Type', y='Total Sales', data=sales_by_area_type, palette='viridis')
plt.title('Total Sales by Area Type (Urban vs. Rural)')
plt.xlabel('Area Type')
plt.ylabel('Total Sales')
plt.show()
```

```
# Plot the average sales per order by Urban vs. Rural
plt.figure(figsize=(10, 6))
sns.barplot(x='Area Type', y='Average Sales per Order', data=sales_by_area_type, palette='viridis')
plt.title('Average Sales per Order by Area Type (Urban vs. Rural)')
plt.xlabel('Area Type')
plt.ylabel('Average Sales per Order')
plt.show()
```

```
# Analyze sales by product category in Urban vs. Rural
category_sales_by_area = df.groupby(['Area Type', 'Category'])['Sales'].sum().reset_index()

plt.figure(figsize=(12, 8))
sns.barplot(x='Category', y='Sales', hue='Area Type', data=category_sales_by_area, palette='muted')
plt.title('Sales by Product Category in Urban vs. Rural Areas')
plt.xlabel('Product Category')
plt.ylabel('Total Sales')
plt.xticks(rotation=45)
plt.show()
```

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

# Step 1: Define the path to your Excel file
file_path = r"D:\SuperStore Sales_cleaned.csv"  # Adjust the path as needed

# Step 2: Read the Excel file
data = pd.read_csv(file_path)

# Step 3: Convert date columns to datetime with dayfirst=True
data['Order Date'] = pd.to_datetime(data['Order Date'], dayfirst=True)
data['Ship Date'] = pd.to_datetime(data['Ship Date'], dayfirst=True)

# Step 4: Calculate shipping time
data['Shipping Time'] = (data['Ship Date'] - data['Order Date']).dt.days

# Step 5: Group by city and calculate average shipping time
avg_shipping_time = data.groupby('City')['Shipping Time'].mean().reset_index()

# Step 6: Sort by average shipping time and get the top 20 cities
top_20_cities = avg_shipping_time.sort_values(by='Shipping Time', ascending=False).head(20)

# Step 7: Plot the results
plt.figure(figsize=(12, 8))
plt.bar(top_20_cities['City'], top_20_cities['Shipping Time'], color='violet')
plt.title('Top 20 Cities with Longest Average Shipping Time')
plt.xlabel('City')
plt.ylabel('Average Shipping Time (days)')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```

```python
import pandas as pd
import matplotlib.pyplot as plt

# Ensure 'Order Date' is in datetime format
df['Order Date'] = pd.to_datetime(df['Order Date'])

# Group by Order Month and Category to get the total sales
df['Order Month'] = df['Order Date'].dt.to_period('M')
category_sales = df.groupby(['Order Month', 'Category'])['Sales'].sum().reset_index()

# Pivot for better visualization
category_sales_pivot = category_sales.pivot(index='Order Month', columns='Category', values='Sales')

# Plotting
category_sales_pivot.plot(kind='line', figsize=(12, 6))
plt.title('Sales of Categories According to Time')
plt.xlabel('Order Month')
plt.ylabel('Total Sales')
plt.xticks(rotation=45)
plt.legend(title='Category')
plt.show()
```

```python
# Group by Order Month and Region to get the total sales
region_sales = df.groupby(['Order Month', 'Region'])['Sales'].sum().reset_index()

# Pivot for better visualization
region_sales_pivot = region_sales.pivot(index='Order Month', columns='Region', values='Sales')

# Plotting
region_sales_pivot.plot(kind='line', figsize=(12, 6))
plt.title('Sales Over Time vs. Region')
plt.xlabel('Order Month')
plt.ylabel('Total Sales')
plt.xticks(rotation=45)
plt.legend(title='Region')
plt.show()
```