# RETAIL SALES ANALYSIS

By: Divyanshu Kumar
(TI_JAD_#12369)

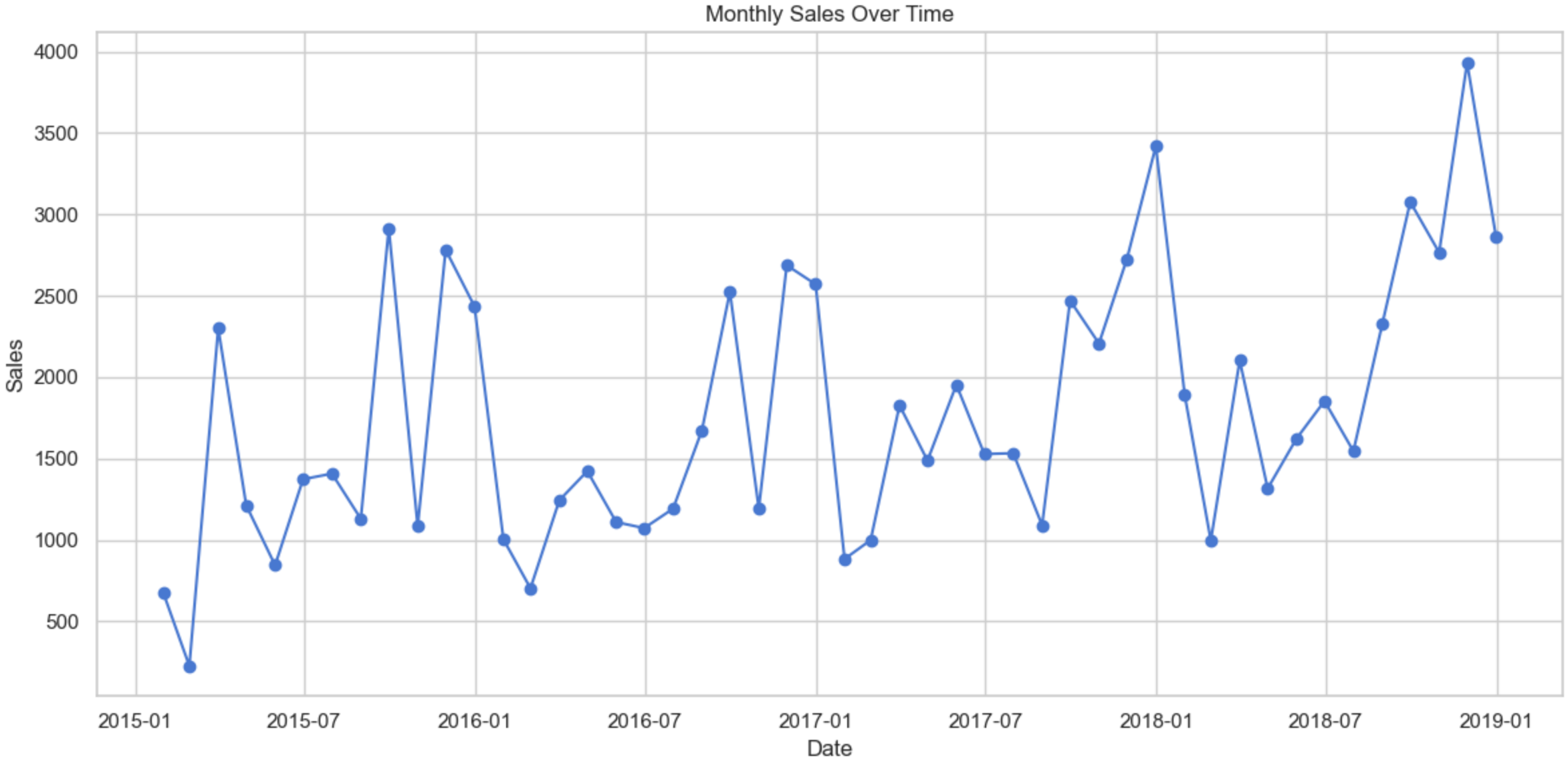
The dataset contains 9,800 records with 18 columns, including details such as Order Date, Ship Date, Ship Mode, Customer Name, Segment, City, State, Postal Code, Region, Product ID, Category, Sub-Category, Product Name, and Sales. The analysis focuses on data cleaning, basic exploration, and visualization, with the following key steps:

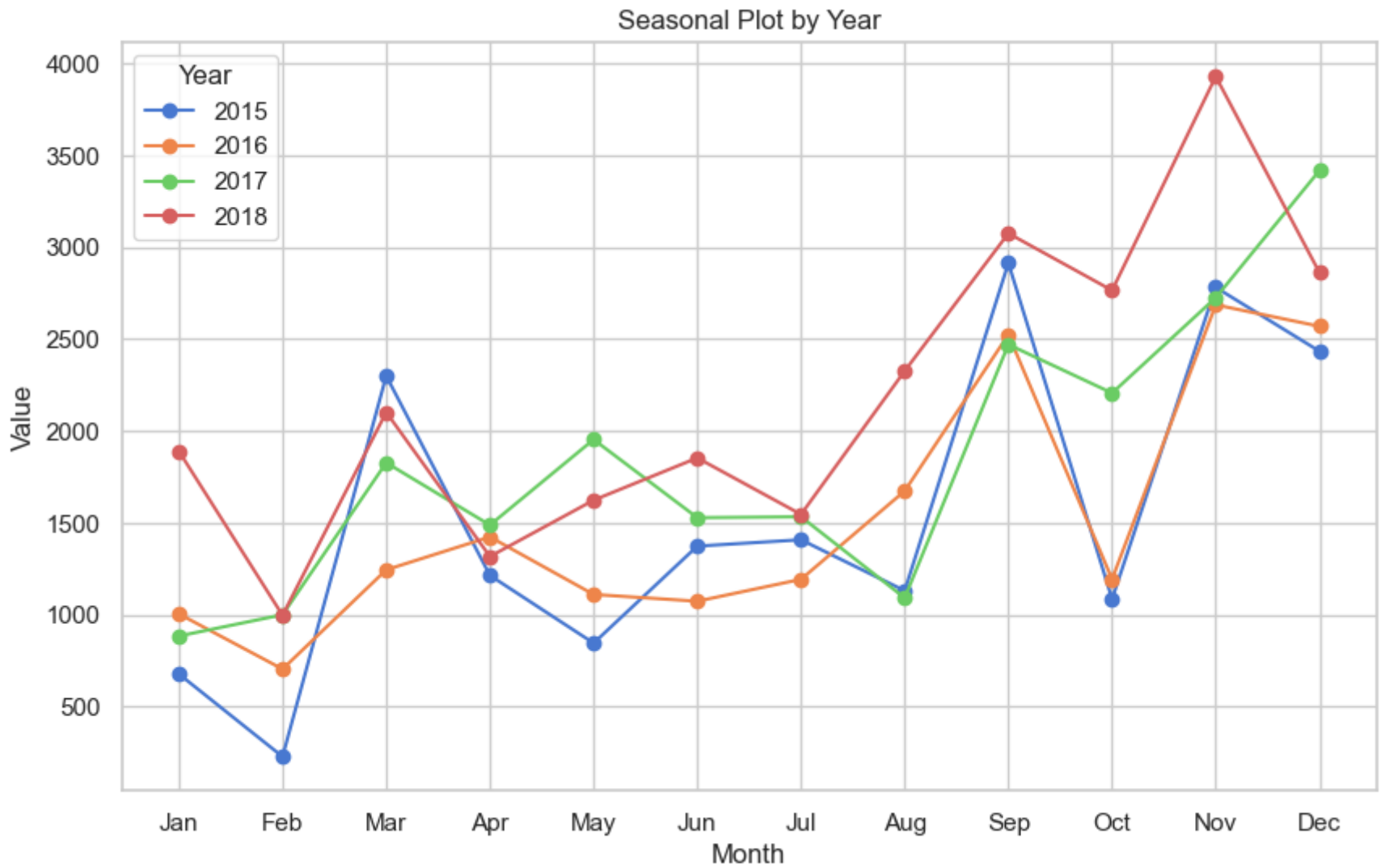
1. **Environment Setup**:
    1. Installed the statsmodels library to resolve the initial ModuleNotFoundError.
    2. Imported necessary libraries: pandas, numpy, matplotlib, seaborn, scipy, itertools, and statsmodels.tsa.seasonal.

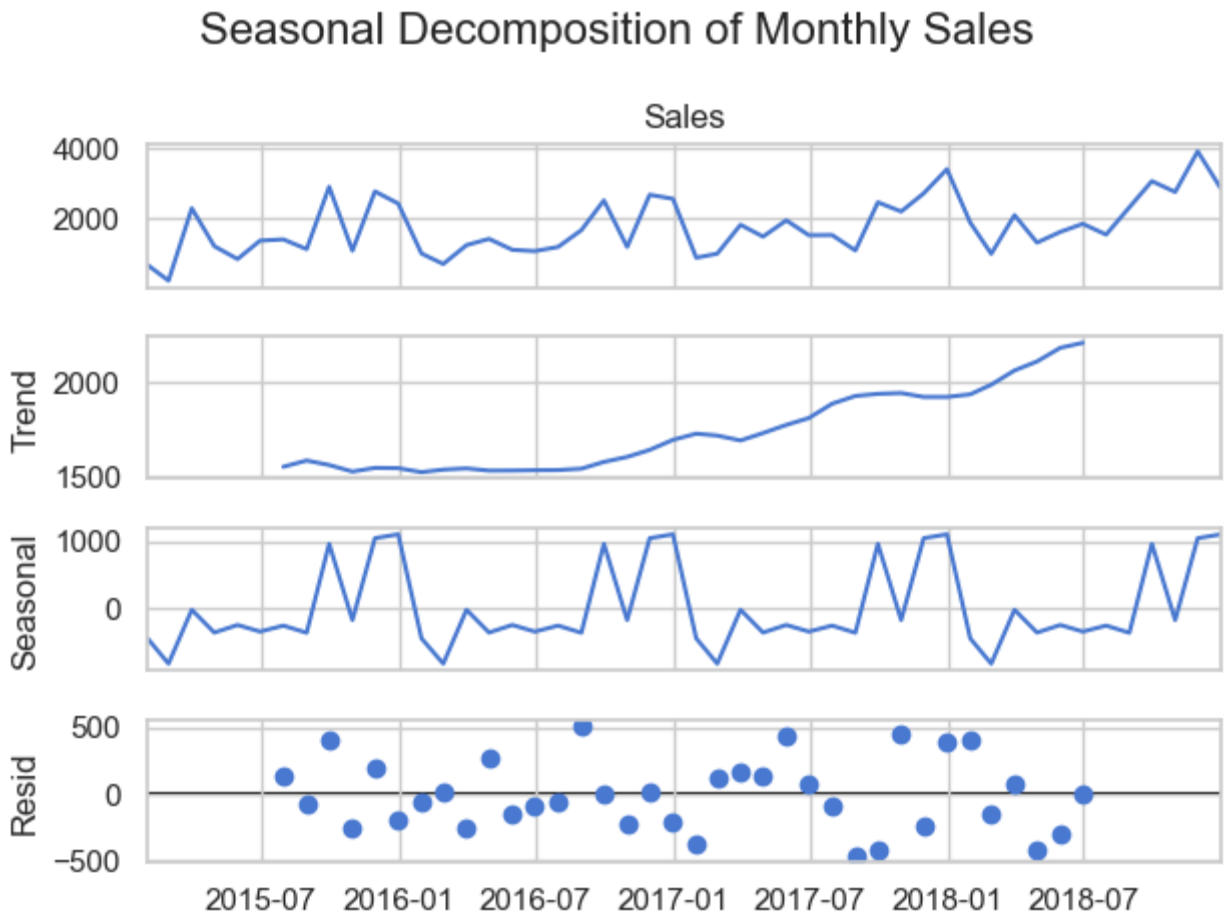
2. **Data Loading and Inspection**:
    1. Loaded the dataset using pd.read_csv with latin1 encoding and parsed Order Date and Ship Date as dates.
    2. Inspected the data using df.head(), df.info(), and df.nunique() to understand structure and unique values.
    3. Found 11 missing values in Postal Code and confirmed data types (mostly object, with Sales as float64 and Row ID as int64).

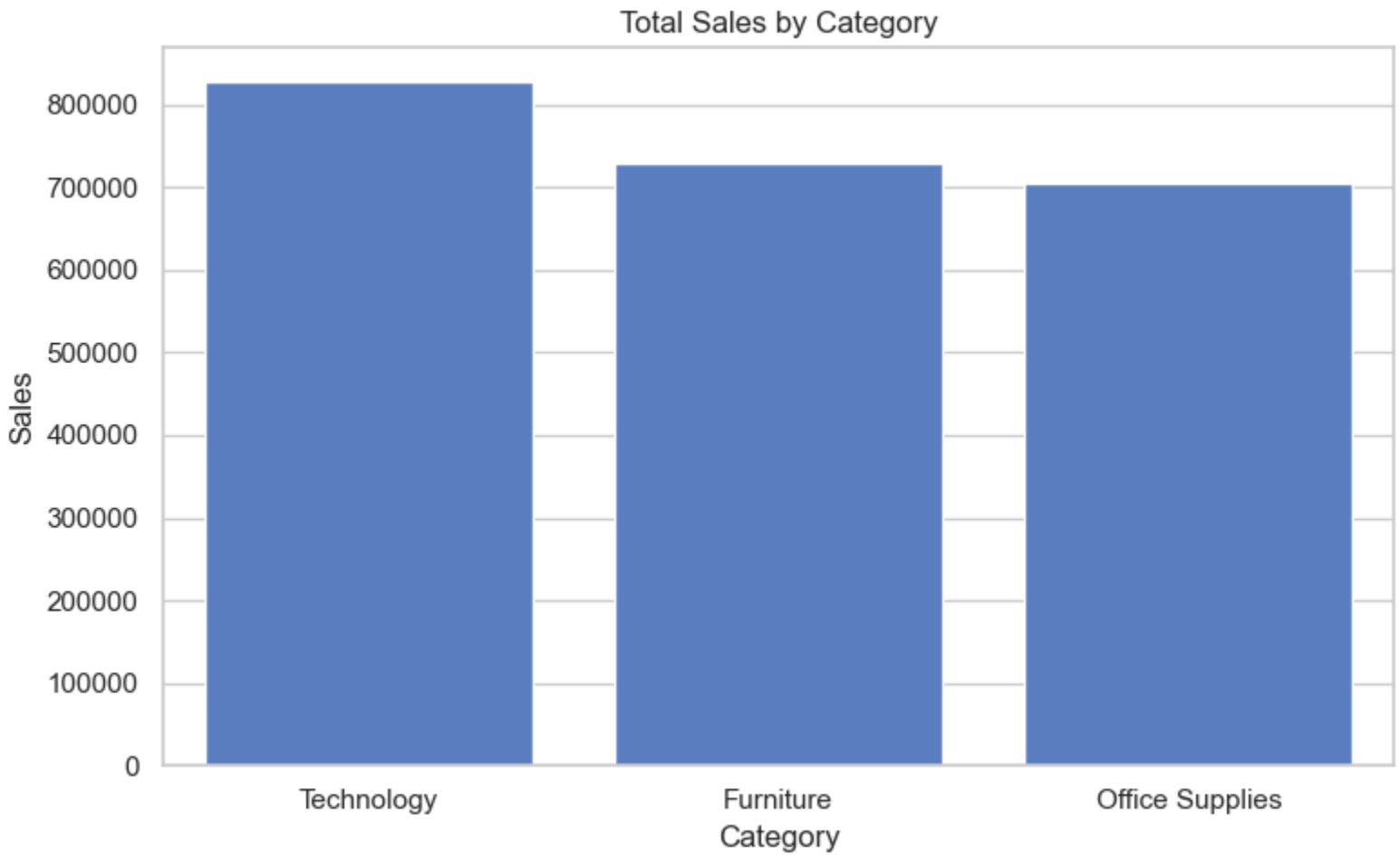
3. **Data Cleaning**:
    1. Dropped unnecessary columns: Row ID, Order ID, Customer ID, Product ID, Customer Name, and Country.
    2. Converted Order Date and Ship Date to datetime format using the correct format (%d/%m/%Y).
    3. Identified missing Postal Code values, all associated with Burlington, Vermont.

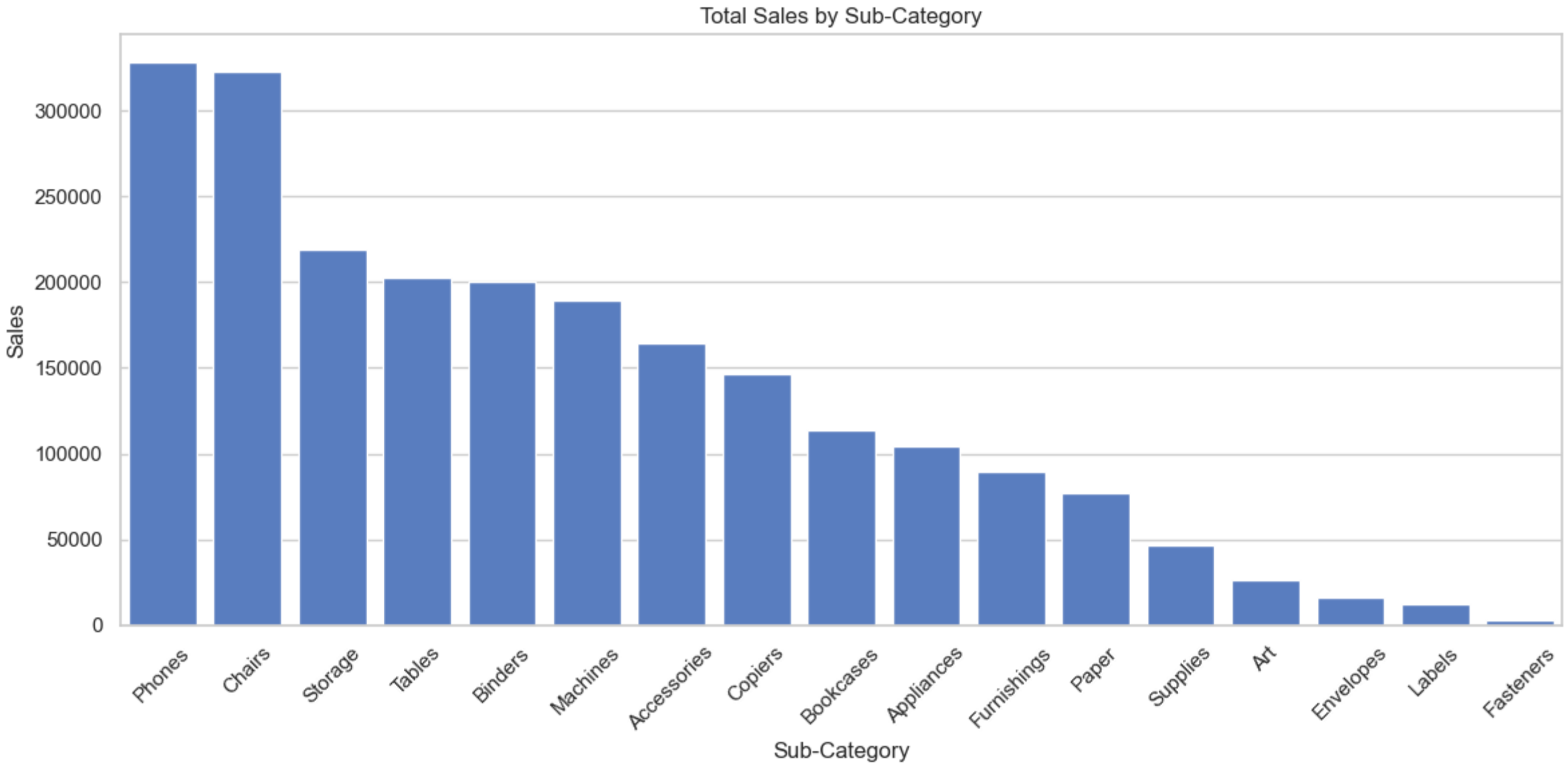
4. **Data Transformation and Visualization**:
    1. Created a Sales_log column using np.log1p(df['Sales']) to handle skewed sales data.
    2. Generated a histogram with a kernel density estimate (KDE) of Sales_log to visualize the distribution of log-transformed sales.
    3. Saved the cleaned DataFrame to Cleaned_Sales.csv.

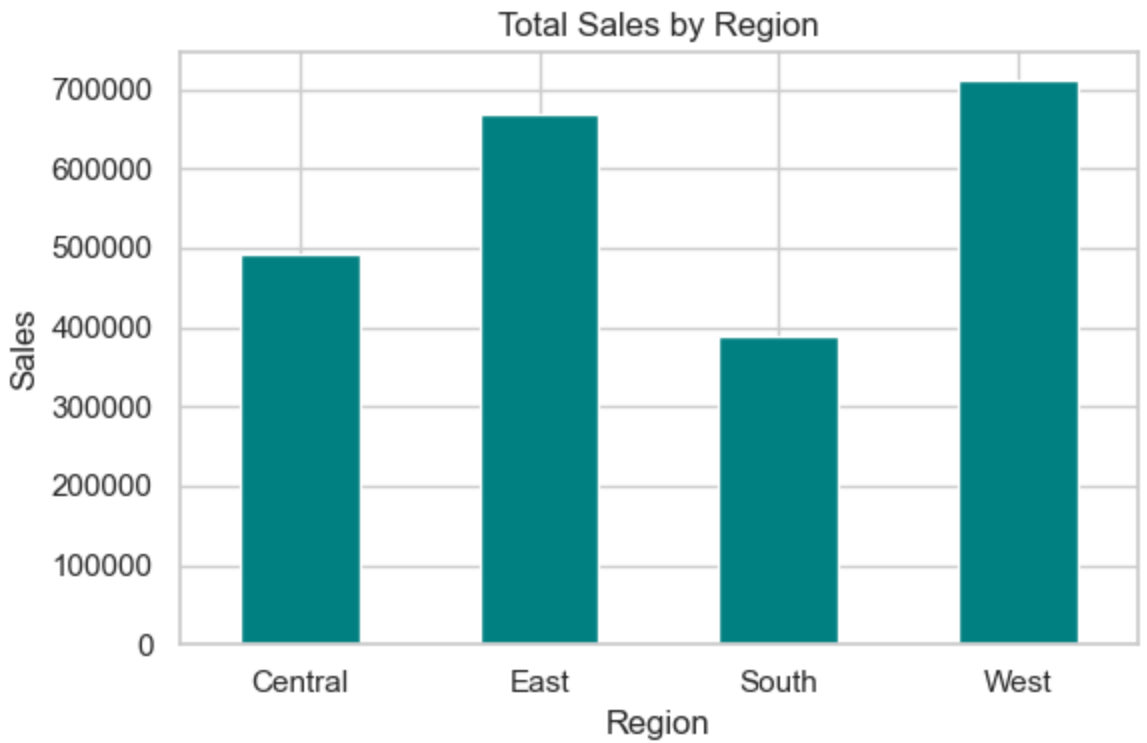

Monthly Sales Over Time
Sales
4000
3500
3000
2500
2000
1500
1000
500
2015-01
2015-07
2016-01
2016-07
2017-01
2017-07
2018-01
2018-07
2019-01
Date

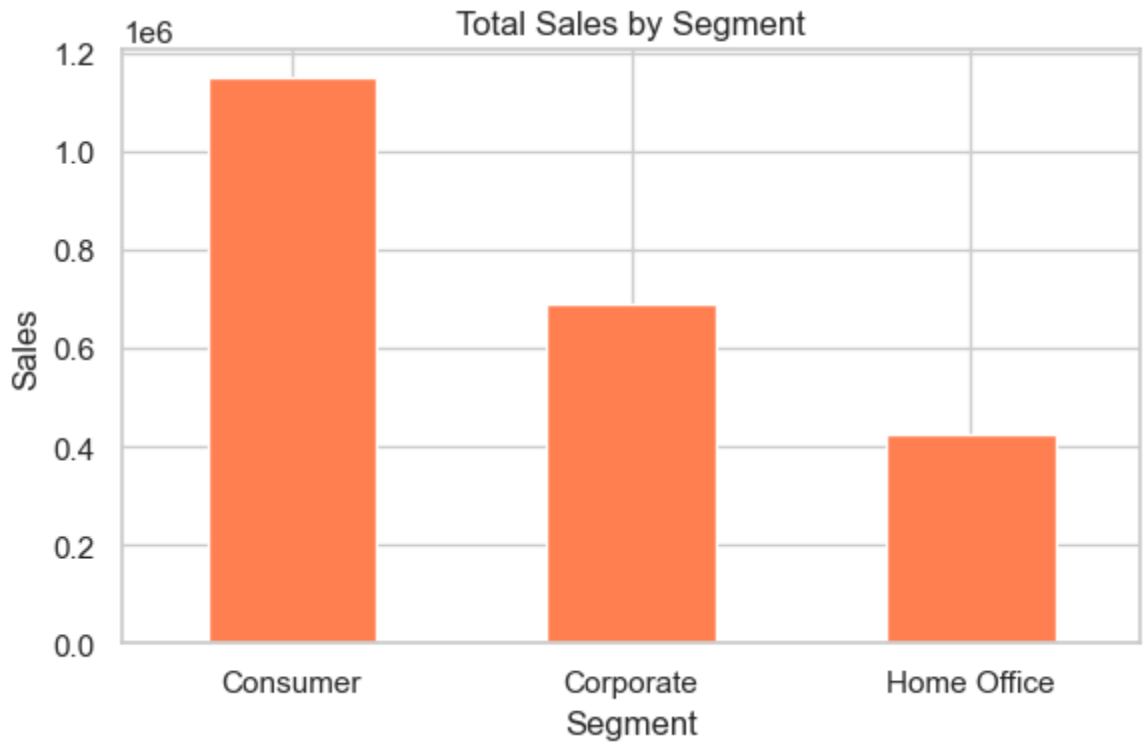

Seasonal Plot by Year
Year
2015
2016
2017
2018
4000
3500
3000
2500
2000
1500
1000
500
Value
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month

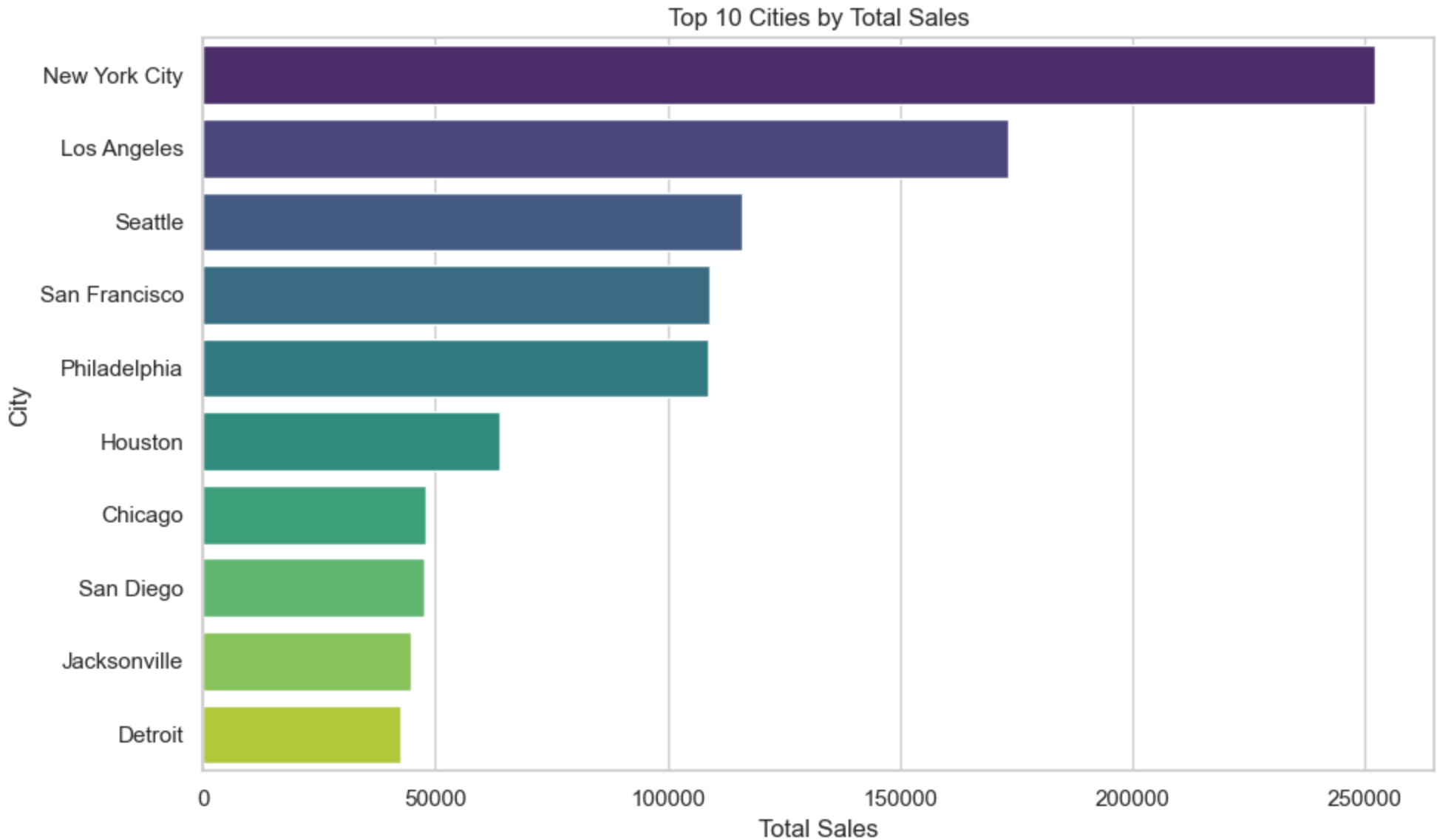

Seasonal Decomposition of Monthly Sales
Sales
4000
2000
Trend
2000
1500
Seasonal
1000
0
Resid
500
0
−500
2015-07
2016-01
2016-07
2017-01
2017-07
2018-01
2018-07

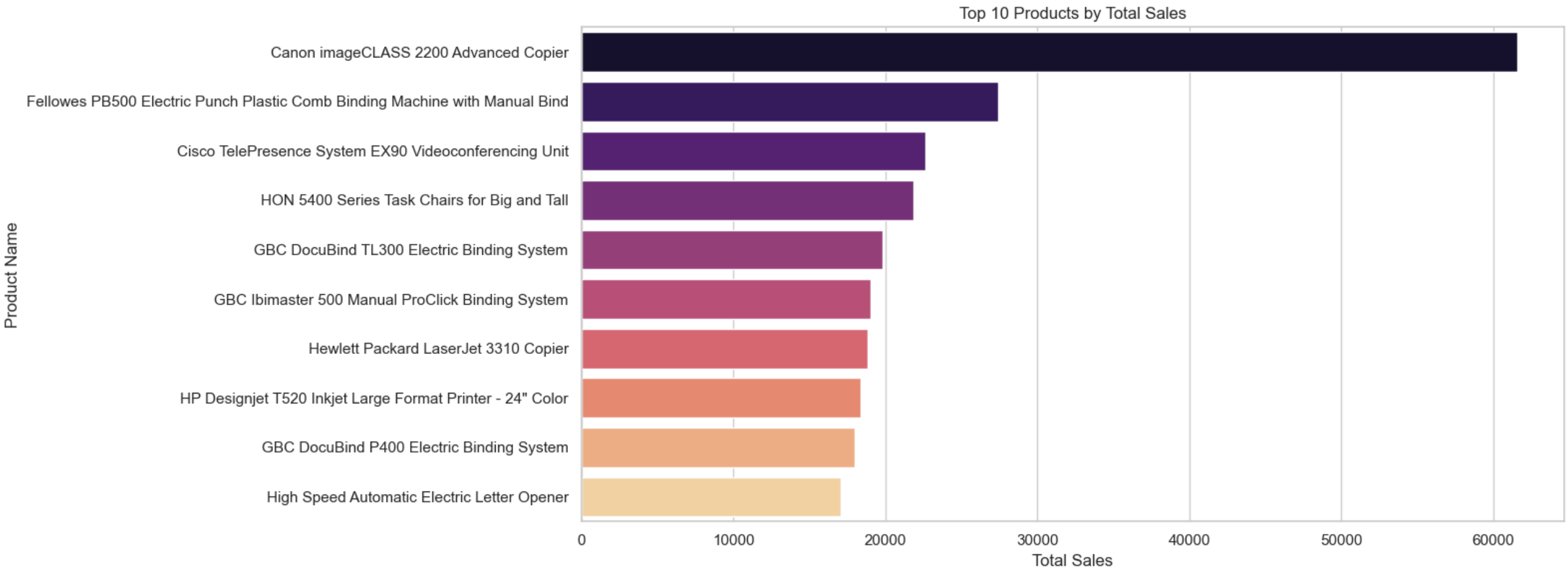

Total Sales by Category
800000
700000
600000
500000
400000
300000
200000
100000
0
Sales
Technology
Furniture
Office Supplies
Category

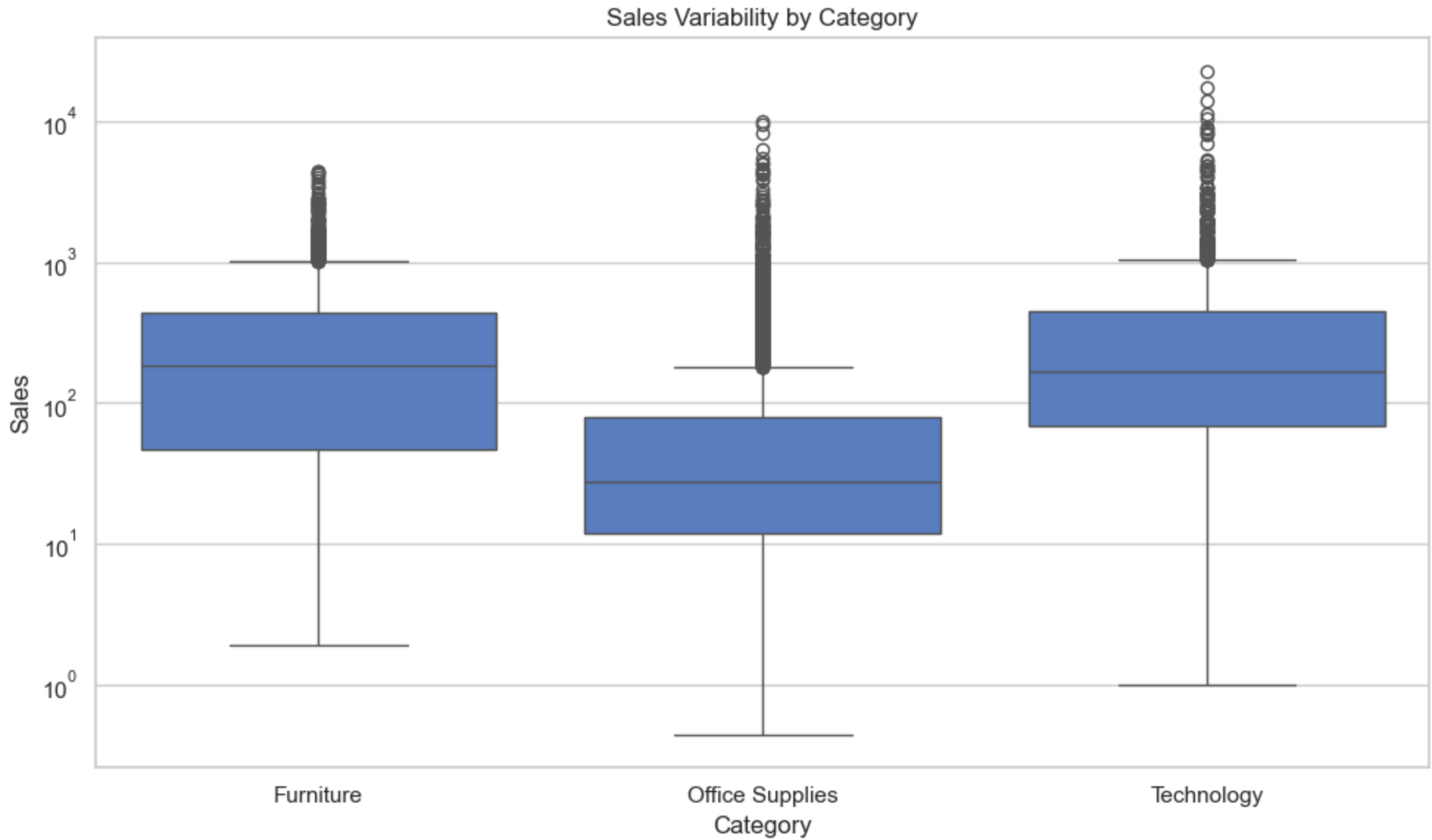

Total Sales by Sub-Category
Sales
300000
250000
200000
150000
100000
50000
0
Phones
Chairs
Storage
Tables
Binders
Machines
Accessories
Copiers
Bookcases
Appliances
Furnishings
Paper
Supplies
Art
Envelopes
Labels
Fasteners
Sub-Category

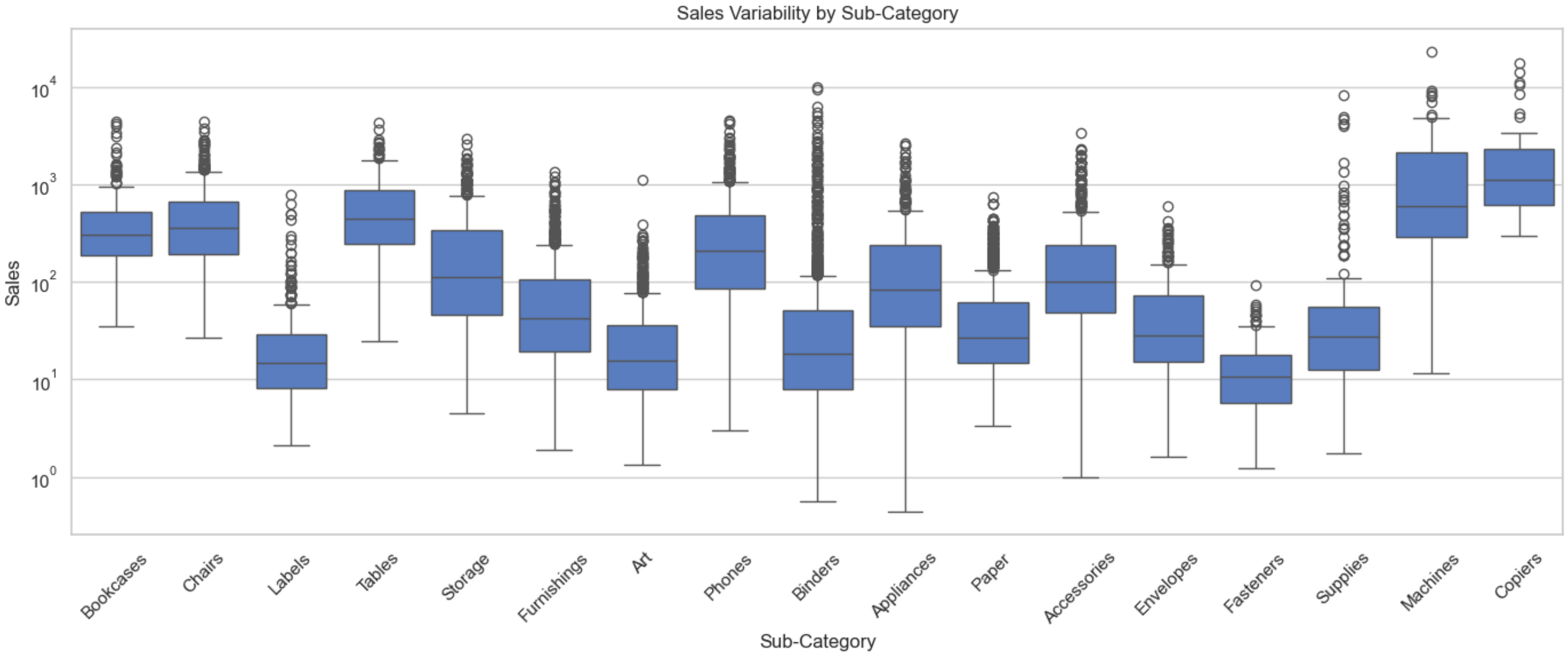

Total Sales by Region
700000
600000
500000
400000
300000
200000
100000
0
Sales
Central
East
South
West
Region

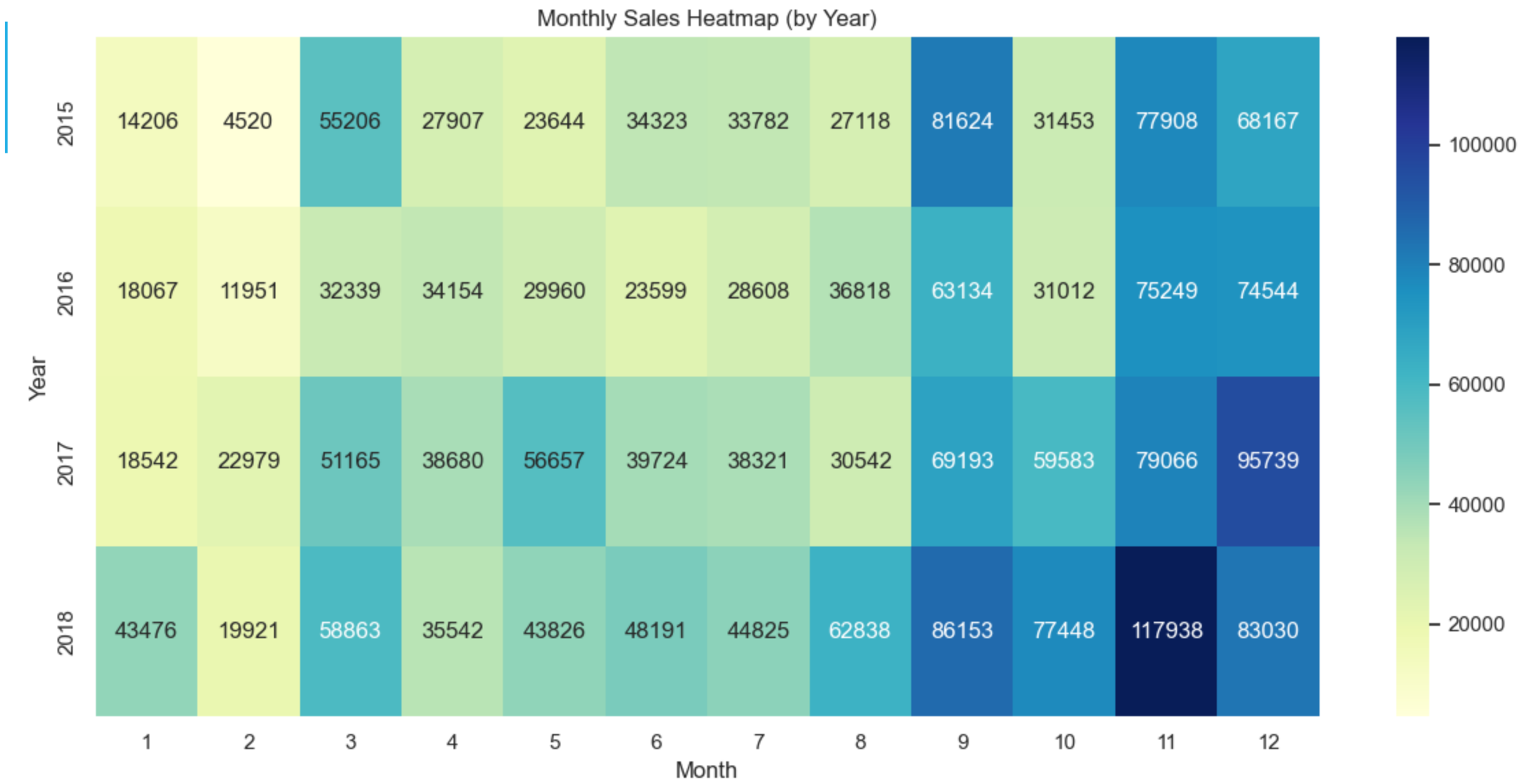

Total Sales by Segment
1e6
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Sales
Consumer
Corporate
Home Office
Segment

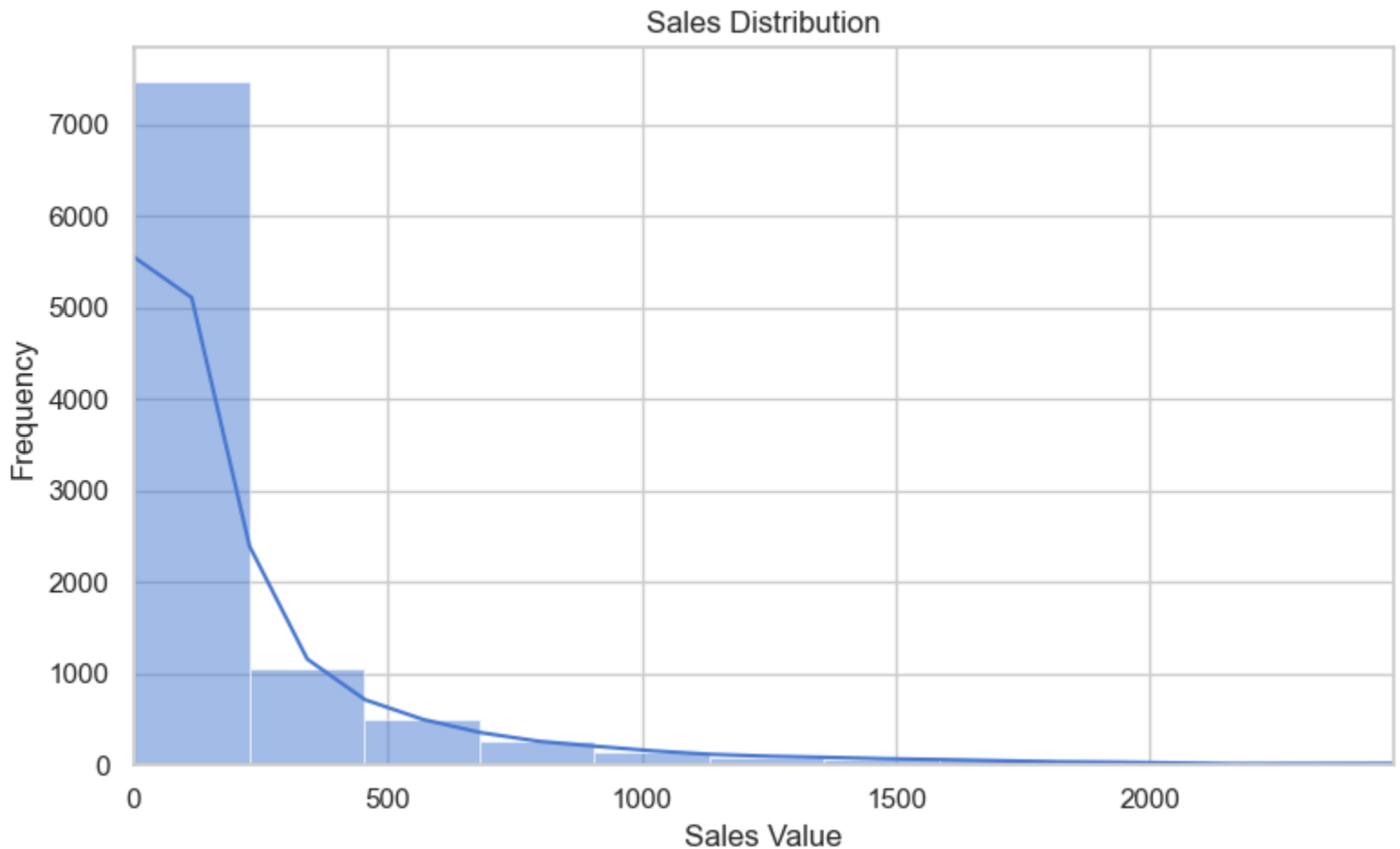

Top 10 Cities by Total Sales
New York City
Los Angeles
Seattle
San Francisco
Philadelphia
Houston
Chicago
San Diego
Jacksonville
Detroit
City
0
50000
100000
150000
200000
250000
Total Sales

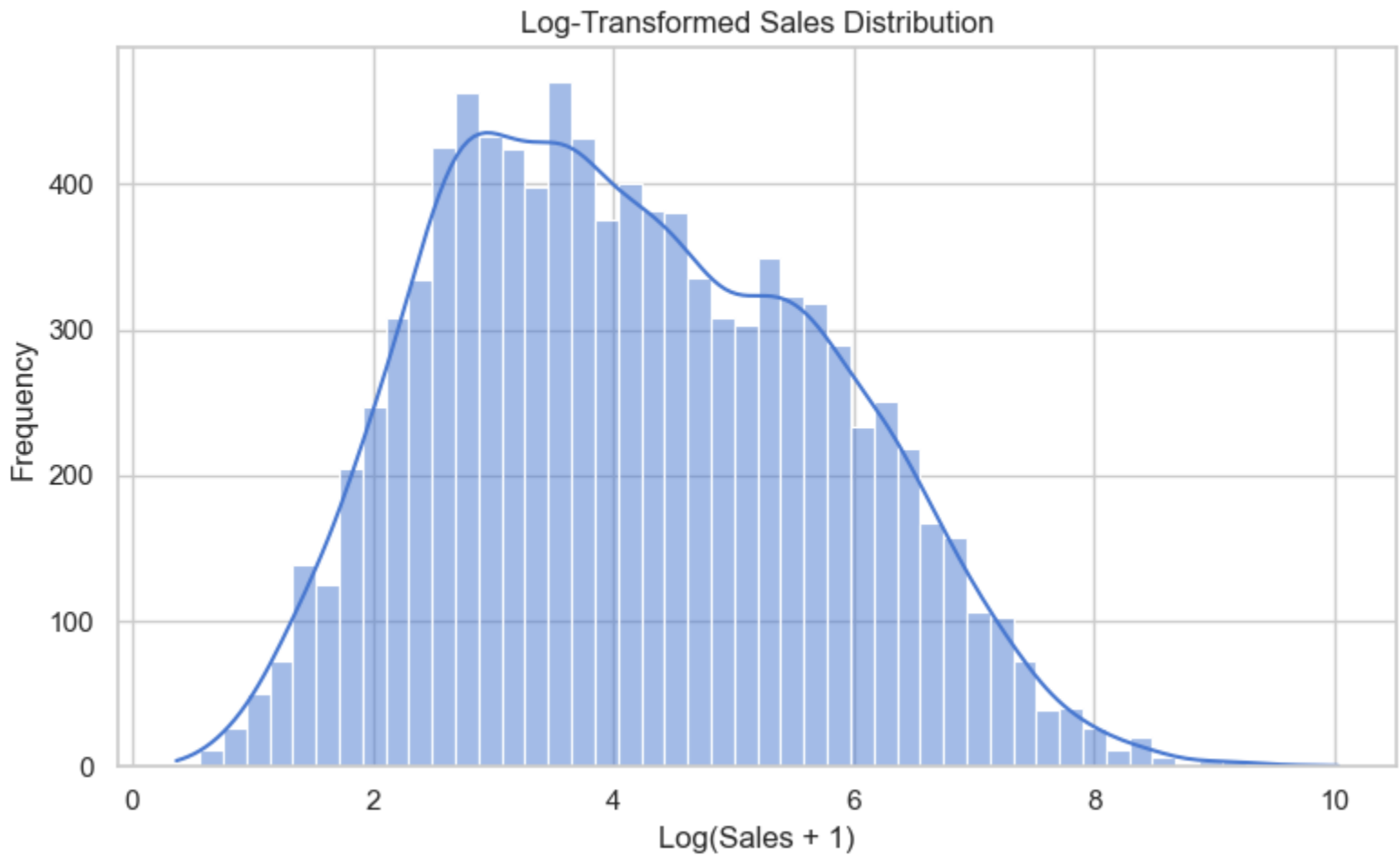

Top 10 Products by Total Sales
Canon imageCLASS 2200 Advanced Copier
Fellowes PB500 Electric Punch Plastic Comb Binding Machine with Manual Bind
Cisco TelePresence System EX90 Videoconferencing Unit
HON 5400 Series Task Chairs for Big and Tall
GBC DocuBind TL300 Electric Binding System
GBC Ibimaster 500 Manual ProClick Binding System
Hewlett Packard LaserJet 3310 Copier
HP Designjet T520 Inkjet Large Format Printer - 24" Color
GBC DocuBind P400 Electric Binding System
High Speed Automatic Electric Letter Opener
Product Name
0
10000
20000
30000
40000
50000
60000
Total Sales


Sales Variability by Category
Sales
$10^4$
$10^3$
$10^2$
$10^1$
$10^0$
Furniture
Office Supplies
Technology
Category


Sales Variability by Sub-Category
Sales
$10^4$
$10^3$
$10^2$
$10^1$
$10^0$
Bookcases
Chairs
Labels
Tables
Storage
Furnishings
Art
Phones
Binders
Appliances
Paper
Accessories
Envelopes
Fasteners
Supplies
Machines
Copiers
Sub-Category


Monthly Sales Heatmap (by Year)
Year
2015
2016
2017
2018
14206 4520 55206 27907 23644 34323 33782 27118 81624 31453 77908 68167
18067 11951 32339 34154 29960 23599 28608 36818 63134 31012 75249 74544
18542 22979 51165 38680 56657 39724 38321 30542 69193 59583 79066 95739
43476 19921 58863 35542 43826 48191 44825 62838 86153 77448 117938 83030
1 2 3 4 5 6 7 8 9 10 11 12
Month
100000
80000
60000
40000
20000


Sales Distribution
Frequency
7000
6000
5000
4000
3000
2000
1000
0
0
500
1000
1500
2000
Sales Value


Log-Transformed Sales Distribution
Frequency
0
100
200
300
400
0
2
4
6
8
10
Log(Sales + 1)