Data Analysis and Machine Learning Implementation

I. Project Overview

The analysis aims to study the key user attributes such as Retailer, Retailer ID, Invoice Date, Region, State, City, Gender Type, Product Category, Price per Unit, Units Sold, Total Sales, Operating Profit, Operating Margin, and Sales Method, using these data points to derive insights into user preferences and behavior.

1. **Retailer** - The retailer allows for efficient data filtering, ensuring that only relevant information pertaining to a particular retailer is processed. Also, enhances the precision, scalability, and versatility of data processing and analysis tasks.

2. **Gender Type** - Analyze sales data based on gender, allowing gender-specific insights into product preferences, buying behavior, and sales trends. Also, this help for data analysis processes enables businesses to gain actionable insights into customer demographics and tailor their strategies accordingly for improved sales performance and customer experience.

3. **Location(Region. State, City)** - Location provide crucial geographical information that aids in understanding the distribution and localization of sales activities. This allows for localized insights into sales performance,

customer demographics, and market trends, which can inform strategic decision-making processes such as store location planning, targeted marketing campaigns, and inventory management.

4. Product Category - Sales data by product type, providing insights into product performance, popularity, and demand. This helps in optimizing inventory management, identifying top-selling products, and guiding marketing strategies.

5. **Price per Unit** - Crucial for calculating total sales revenue, understanding pricing strategies, and assessing product profitability. It helps segment data to analyze pricing trends and compare prices across products like Variations in prices for men's and women's products can reflect differences in design, production costs, market demand, and retailer pricing strategies, providing valuable insights for inventory management and marketing decisions.

6. Units Sold - It helps quantify sales volume, identify best-selling products, and analyze demand trends. Examining units sold, businesses can assess product performance, optimize inventory levels, and forecast future sales.

7. **Total Sales** - It represents the revenue generated from each transaction, providing a direct measure of financial performance. Analyzing total sales

helps identify high-revenue products and peak sales periods, and assess the effectiveness of pricing strategies.

8. Operating Profit - It shows the profit made from sales after accounting for operating expenses. Analyzing operating profit helps businesses understand which products or categories are most profitable, identify cost-saving opportunities, and evaluate overall financial health.

9. Operating Margin - It indicates the efficiency of a company's operations by showing the percentage of revenue that remains as profit after covering operating expenses. Analyzing operating margin helps assess profitability across different products, categories, or time periods, highlighting areas with higher financial efficiency.

10. Sales Method - It identifies the channel through which sales were made, such as outlets or online stores. Analyzing the sales method helps businesses understand the performance of different sales channels, customer preferences, and the effectiveness of various marketing strategies.

By analyzing this attributes, Addidas can gain valuable insights into customer demographics, geographic performance, and product popularity. This analysis will also help in understanding sales trends, profits, and the effectiveness of different sales channels, companies can make smarter decisions about where to invest resources.

II. Libraries and Data Handling

Libraries Used:

Pandas for Data Manipulation and analysis, Seaborn for Statistical data visualization based on matplotlib, Matplotlib for Plotting and visualization, Scikit-learn for Machine learning, including model selection, training, and evaluation, along with preprocessing utilities, and Statsmodel for Statistical modeling and hypothesis testing, particularly for time series analysis.

1. **Pandas** - This library is for data manipulation and analysis. It offer tools and structures for managing numerical tables and time series, making it perfect for analyzing and processing large datasets like the Addidas Sales Analysis.

2. **Seaborn** - This library is for statistical data visualization, based on Matplotlib, that simplifies the process of creating complex visualizations, such as heatmaps, time series plots, and categorical plots, by providing high-level functions and built-in themes and color palettes. This makes it useful for exploring and understanding data patterns and relationships.

3. Matplotlib - This library is for creating static, animated, and interactive visualizations. It offers plotting functions for generating various types of plots, including line plots, scatter plots, histograms, bar charts, and more. This is used in scientific computing, data analysis, and visualization tasks due to its flexibility and extensive capabilities.

4. Scikit-learn - This library is for machine learning. It offers various machine learning algorithms for classification, regression, clustering, dimensionality reduction, and more. Also, this includes utilities for data preprocessing, model evaluation, and model selection.

5. Statsmodel - This library is for estimating and interpreting statistical models. It provides a wide range of tools for conducting statistical analysis, hypothesis testing, and econometric modeling. It also offers functionality for exploring data, conducting statistical tests, and visualizing results.

Data Loading and Preprocessing

- **Data Loading**: Loading data from a CSV file into python involves libraries such as pandas or Python's built-in csv module. Using **import pandas as pd** and then use the **pd.read_csv()** function for it to read the CSV file into a pandas as a DataFrame, enabling data manipulation operations such as filtering, sorting, aggregation, and visualization.

Data Cleaning and Preprocessing

- **Handling Dates** – The "Invoice Date" column in our Dataset was converted to datetime format using '**pd.to_datetime()**', and then set as the index for time series analysis.

- **Handling Missing Values** – Missing values were identified using '**isnull()**' and the count of missing values was obtained using '**sum()**'. A heatmap visualization **'sns.heatmap()'** was created to visualize the missing data pattern.

- **Descriptive Statistics** – Descriptive statistics of the dataset were generated using **'describe()'** to gain insights into the data distribution and identify potential outliers.

- **Categorical Data** – Categorical variables like "Region", "Product Category", "Sales Method", and "Gender Type" were analyzed using value counts **('value_counts()')** and visualized using bar plots **('plt.bar()' or 'sns.countplot()')** to understand the distribution and trends within each category.

- **Data Transformation** – For machine learning implementation, data was prepared by dropping irrelevant columns ("Retailer ID", "State", "City"), creating a binary target variable ("High_Sales") based on a threshold, and generating dummy variables for categorical features using **‘pd.get_dummies()’**.
- **Standardization** – Data standardization was performed using **‘StandardScaler()’** to scale numerical features before training the logistic regression model.

The structure for any Python-based data analysis workflow is laid by these steps, providing a structured approach to comprehend and visualize user data. This is to ensure the dataset's structure, completeness, and compatibility with analysis and modeling techniques, thereby ensuring its readiness for advanced analysis and visualizations.

III. Data Analysis Techniques

Descriptive Statistics: Summary statistic such as mean, median, standard deviation, minimum, and maximum values used to understand the distribution of data. These statistics provide a concise overview of the central tendency, variability, and shape of the dataset, facilitating a better understanding of its characteristics and informing subsequent analyses and decision-making processes. Here Here's how they help in the context of Adidas Sales Analysis:

- **Mean** - The mean provides the average sales value, helping to gauge the overall performance of Adidas products across different regions, product categories, or sales methods. It offers a central reference point for assessing sales trends and identifying areas of strength or weakness.
- **Median**- The median represents the middle value of the sales data when arranged in ascending order. It is less affected by extreme values compared to the mean and provides a robust measure of central tendency. The median helps to understand the typical or typical sales performance, especially in datasets with outliner.

- **Standard Deviation** – The standard deviation measures the dispersion or spread of sales values around the mean. A higher standard deviation indicates greater variability in sales, which can highlight regions, product categories, or sales methods with inconsistent performance. This helps in assessing the reliability and predictability of sales data.

- **Minimum and Maximum Values** – The minimum and maximum sales values indicate the lowest and highest recorded sales figures, respectively. They provide insights into the range of sales performance observed in the dataset. Identifying the minimum and maximum values helps in understanding the breadth of sales data and identifying outliers or exceptionally high-performing regions, products, or sales methods.

Data Visualization: Various plots and charts are used to visualize the distribution of users by region, product category, sales method, and retailer. These visualizations provide a clear representation of the frequency or count of users in each category, enabling easy comparison and interpretation of distribution patterns. Here's how various types of plots are employed in Adidas Sales Analysis:

- **Bar Plots** – used to visualize the distribution of categorical variables such as state, city, region, product category, sales method, and retailer. These plots provide a clear graphical representation of the frequency or count of each category, enabling easy comparison and interpretation.

- **Pie Charts** – used to illustrate the distribution of categorical variables like retailer, region, and product category as proportions of the whole. They provide a visual representation of the relative contribution of each category to the total, making it easier to understand the distribution patterns.

- **Heatmaps** – used to visualize missing values in the dataset. It provides a graphical representation of the presence or absence of missing values across different columns, allowing for easy identification of missing data patterns.

- **Countplots** – used to visualize the distribution of categorical variables like gender type and their relationships with other categorical variables such as region, sales method, and product category. These plots help in understanding the frequency of different categories and their associations.

- **Line Plots** – used to visualize time series data, specifically monthly total sales over time. These plots show the trend in sales over months, helping to identify patterns, seasonality, and trends in Adidas sales data.

These techniques are fundamental for making informed decisions based on user data. Descriptive statistics provide the numerical background necessary to understand the data at a basic level, while visualization techniques help bring this data to life, making it easier for stakeholders to digest and make strategic decisions based on these insights.

IV. Key Findings

Major Findings: These findings from our analysis offer valuable guidance for shaping business decisions and strategies, including strategic resource allocation, targeted marketing efforts, optimized sales strategies, and enhanced competitive positioning.

- **Regional Performance** – Certain regions show higher sales volumes compared to others, indicating geographical variations in consumer preferences and market demand. Understanding these regional trends can help in allocating resources effectively, targeting marketing efforts, and optimizing inventory management strategies based on regional demand.

- **Product Category Preferences** – The analysis reveals differences in sales performance across product categories, with some categories experiencing higher demand than others. Identifying these preferences allows Adidas to focus on popular categories, invest in product development, and tailor marketing campaigns to capitalize on consumer preferences.
- **Sales Method Effectiveness** – The analysis highlights the effectiveness of different sales methods in driving sales. By understanding which sales methods are most successful, Adidas can allocate resources efficiently, optimize its sales channels, and enhance its overall sales strategy to maximize revenue generation.
- **Trends Over Time** – Time series analysis reveals sales trends over months or years, including seasonality patterns and overall sales trajectory. Recognizing these trends enables Adidas to forecast future sales, plan inventory levels, and implement targeted marketing initiatives to capitalize on peak seasons and mitigate seasonal fluctuations.

Business Impact: These findings have significant implications for shaping business decisions and strategies at Adidas:

- **Marketing Strategies –** *Insights into regional performance and product category preferences can inform targeted marketing campaigns tailored to specific regions and consumer segments, maximizing marketing ROI and driving sales growth.*
- **Inventory Management –** *Understanding sales trends and seasonality patterns helps optimize inventory levels, ensuring that popular products are adequately stocked while minimizing excess inventory costs.*
- **Sales Channel Optimization –** *Identifying effective sales methods allows Adidas to allocate resources strategically, invest in high-performing sales channels, and optimize its sales strategy to reach target customers more effectively.*
- **Forecasting and Planning –** *Utilizing time series analysis for sales forecasting enables Adidas to anticipate future demand, plan production schedules, and make informed business decisions to meet customer needs efficiently.*

V. Advanced Analysis

- **Geographical Insights** – By examining regional performance differences, advanced geographical analysis provides insights into consumer preferences and market demand in various geographic areas. By examining sales data on a geographical level, Adidas gains valuable insights into regional market dynamics, enabling targeted marketing efforts, strategic resource allocation, and market expansion opportunities.
- **Temporal Trends** – Time series analysis is utilized to uncover temporal trends in Adidas sales data, including seasonal fluctuations and overall sales trajectories over months or years. By analyzing sales patterns over time, Adidas can identify seasonal trends, peak sales periods, and fluctuations in consumer demand. This enables proactive inventory management, accurate sales forecasting, and strategic planning to capitalize on peak seasons and mitigate seasonal variations effectively.

The comprehension of seasonal variations in Adidas sales data and larger market dynamics is helped by these precise analytical techniques. Adidas may increase its competitive standing in the changing sportswear industry by optimizing sales channels, tailoring business strategies, and gaining this insights into temporal and geographical patterns.

VI. Machine Learning Implementation

Logistic Regression Model

Logistic Regression's is a widely used statistical method for binary classification problems. For our analysis of predicting high sales in the Adidas dataset, logistic regression is a perfect fit because of its interpretability, computing efficiency, robustness, and suitability for binary classification.

Preparing the Data for Logistic Regression

Data Selection and Data Cleaning – To effectively prepare the data for analysis, we first inspect the dataset's shape, column names, and data types. We identify and count missing values, then visualize the missing data pattern to guide our cleaning strategy.

We clean the data by removing irrelevant columns such as "Retailer ID", "State", and "City". These columns are not directly relevant to our analysis of sales performance. Next, we create a new binary target variable, High_Sales, based on the median of Total Sales. Sales values above the median are labeled as 1 (high sales), and those below or equal to the median are labeled as 0.

For categorical variables such as "Gender Type", "Region", "Product Category", and "Sales Method", we use one-hot encoding to convert these categories into numerical values. This transformation is essential for the Logistic Regression model, as it cannot handle categorical data directly.

Feature Scaling - Feature scaling is necessary to standardize the values of the features. We split the dataset into training and testing sets using a 70-30 split. Standardization ensures that all feature values are on a similar scale, which is important for the Logistic Regression model's effectiveness. We apply standard scaling to the training set and use the same scaling parameters to transform the testing set.

Building the Logistic Regression Model

Model Training

- **Splitting Data** - After data preparation, we split the dataset into training and testing sets. The training set, comprising 70% of the data, is used to train the model, while the testing set, containing the remaining 30%, is reserved for evaluation.

- **Model Fitting** - With the training set prepared, we fit the Logistic Regression model to the data. During this process, the model learns the relationship between the predictor variables (features) and the binary target variable (high sales). This involves estimating the coefficients of the model that best describe the relationship.

Model Evaluation:

- **Performance Metrics** - After the model is trained, we evaluate its performance using various metrics. These metrics include accuracy, which measures the proportion of correctly predicted outcomes, and a confusion matrix, which provides a breakdown of correct and incorrect predictions. Also, a detailed classification report offers insights into the model's precision, recall, and F1-score for each class.
- **Residual Analysis** - Residual analysis involves examining the discrepancies between the predicted and actual values. By analyzing the residuals, we can identify any patterns or trends that the model may have missed. This analysis helps us assess the model's overall fit and identify areas for improvement.

VII. Visual Insights

- **Bar Charts, Pie Charts, Heatmaps**: Usage and insights these visuals provide.
- **Value Counts for State** – This bar chart displays the distribution of sales data across different states.
- **Value Counts for City** – The bar chart illustrates the distribution of sales across various cities.
- **Value Counts for Region** – This count plot showcases the distribution of sales across different regions.
- **Value Counts for Product Category** – This count plot presents the distribution of sales across different product categories.
- **Value Counts for Sales Method** – This count plot illustrates the distribution of sales across various sales methods.
- **Value Counts for Retailer** – The count plot displays the distribution of sales across different retailers.
- **Distribution of Retailer** – The pie chart illustrates the distribution of sales among different retailers.

- **Distribution of Region** – The pie chart represents the distribution of sales across different regions.
- **Distribution of Product Category** – This pie chart represents the distribution of sales across different product categories.
- **Distribution of Sales Method** – This pie chart show the distribution of of sales across different sales method.

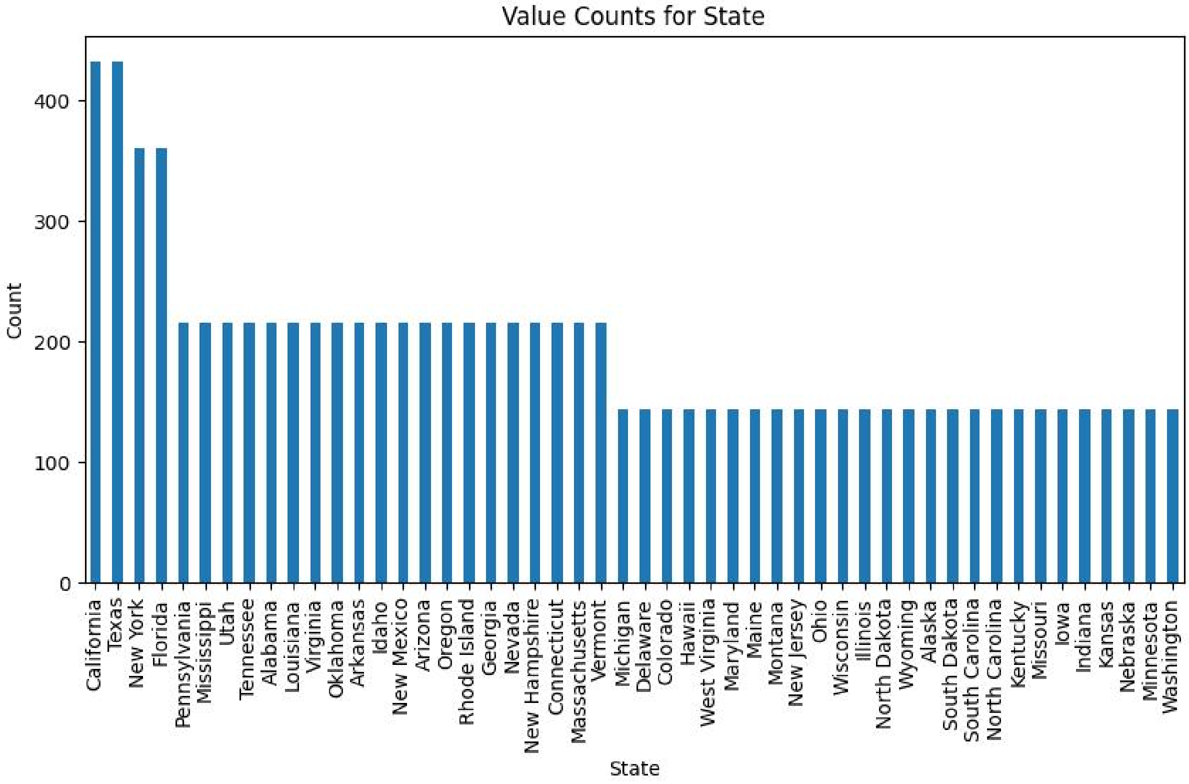

Figure 1.0 The bar chart shows the count of sales in each state. his information can help identify regions with high or low sales volumes, allowing for targeted marketing or resource allocation strategies.

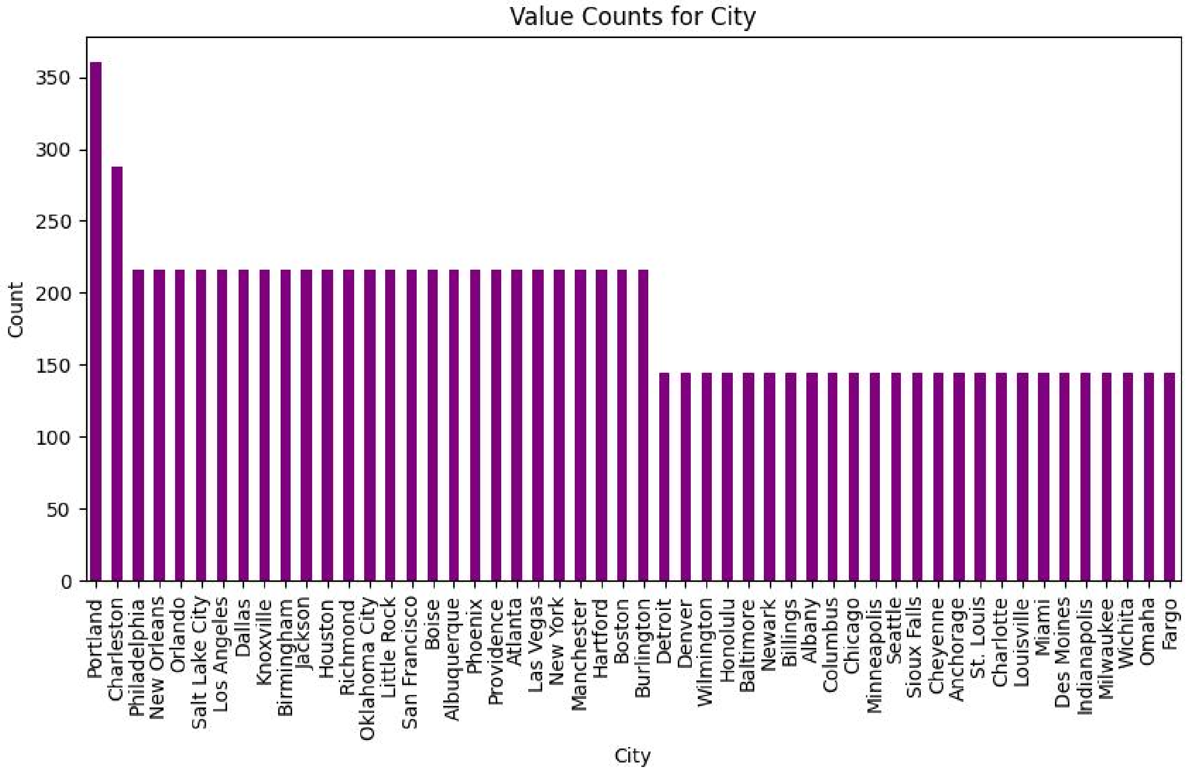

Figure 2.0 The bar chart shows the count of sales in each city. This information can inform decisions related to store locations, inventory management, and promotional activities.

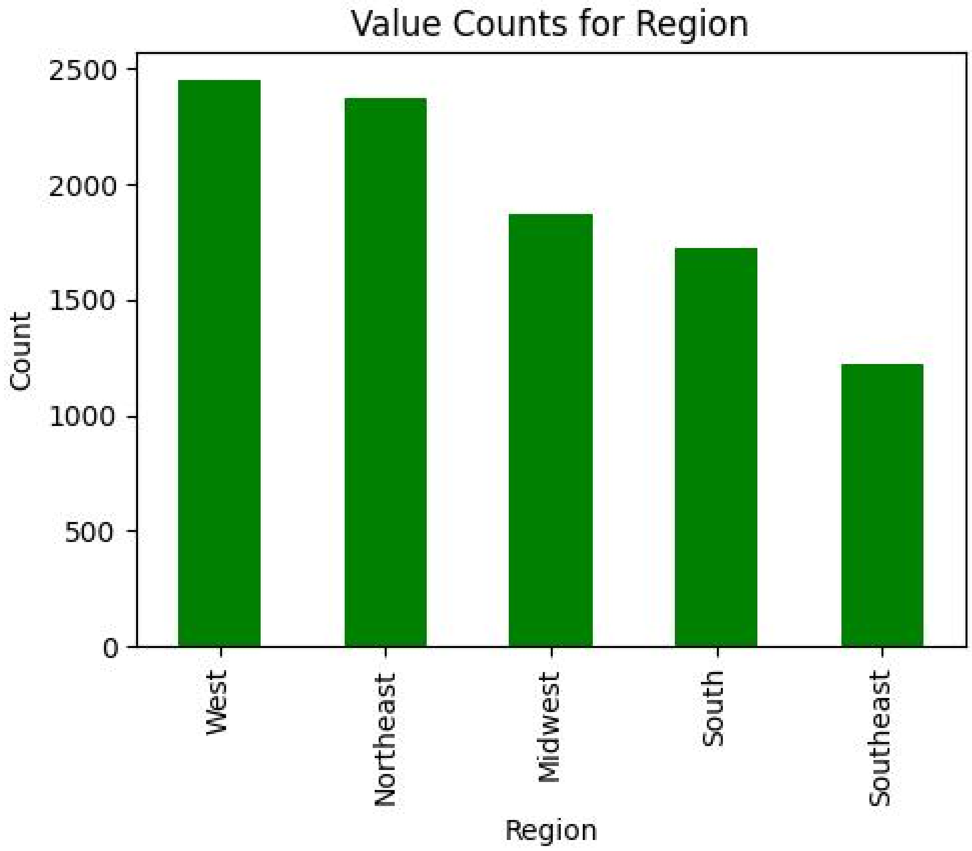

Figure 3.0 The bar chart shows the count of sales in each region. This insight enables businesses to tailor marketing campaigns and product offerings to specific regional preferences and trends.

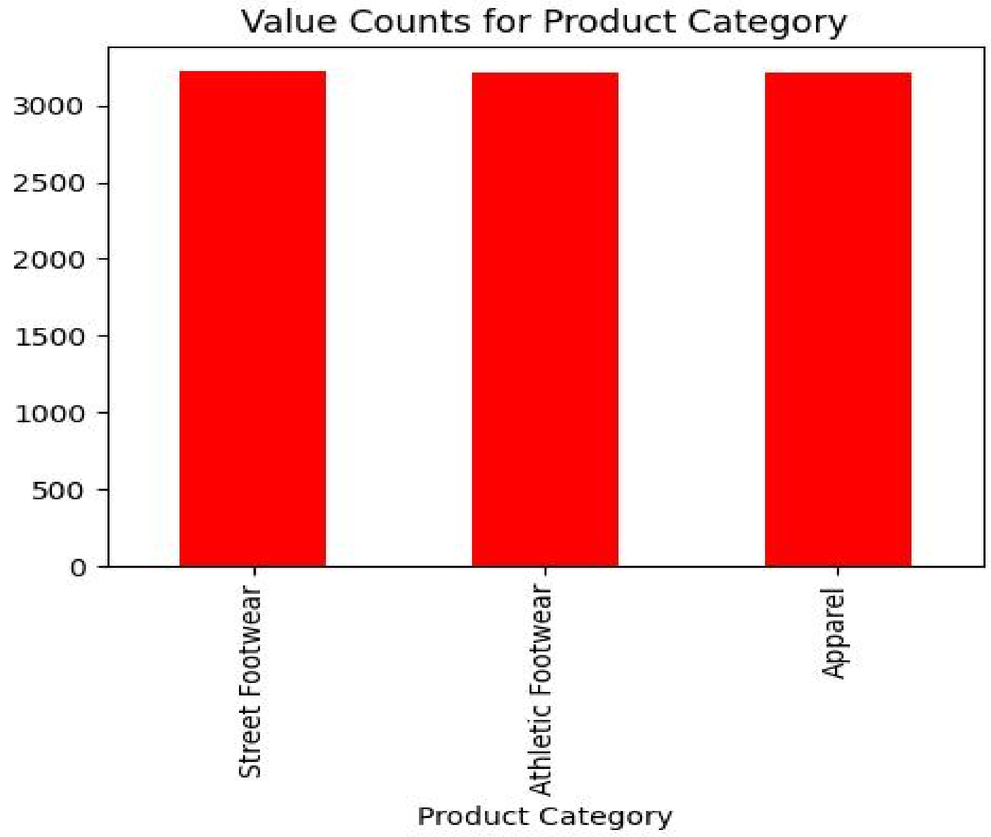

Figure 4.0 The bar chart shows the popularity of each product category. This insights allows businesses to prioritize marketing efforts, optimize inventory management, and identify opportunities for product diversification or expansion.

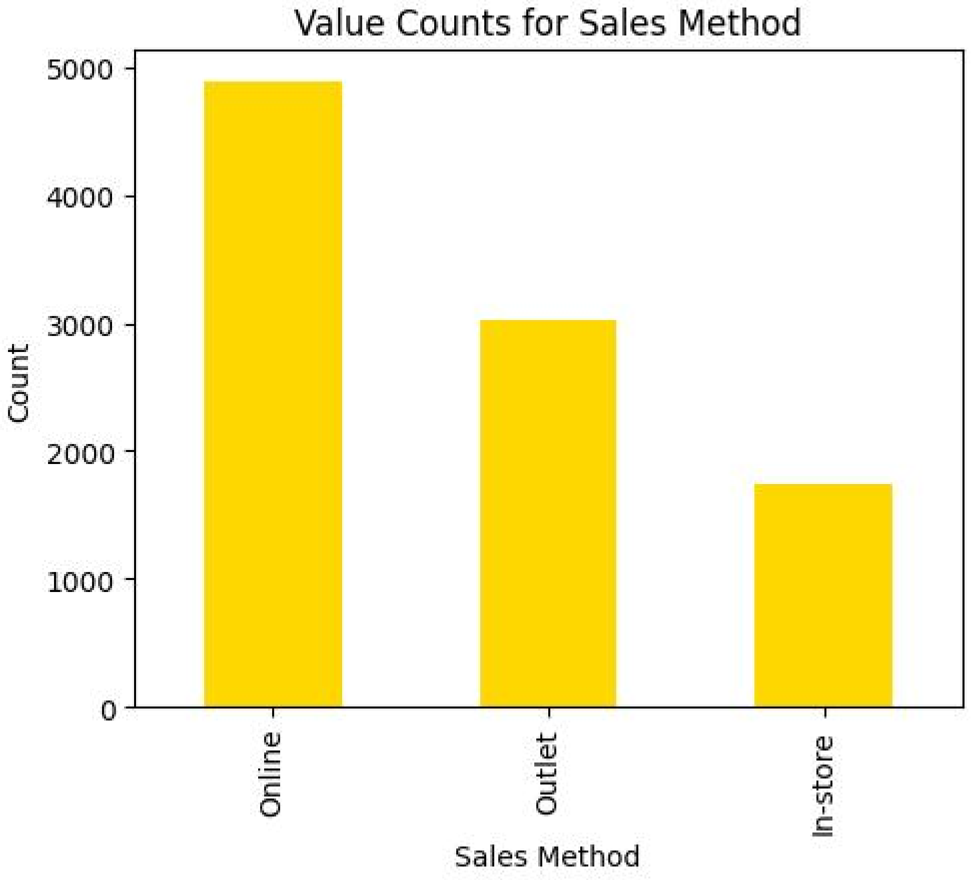

Figure 5.0 The bar chart shows the count of sales for each method. This insight informs strategic decisions related to sales and distribution strategies.

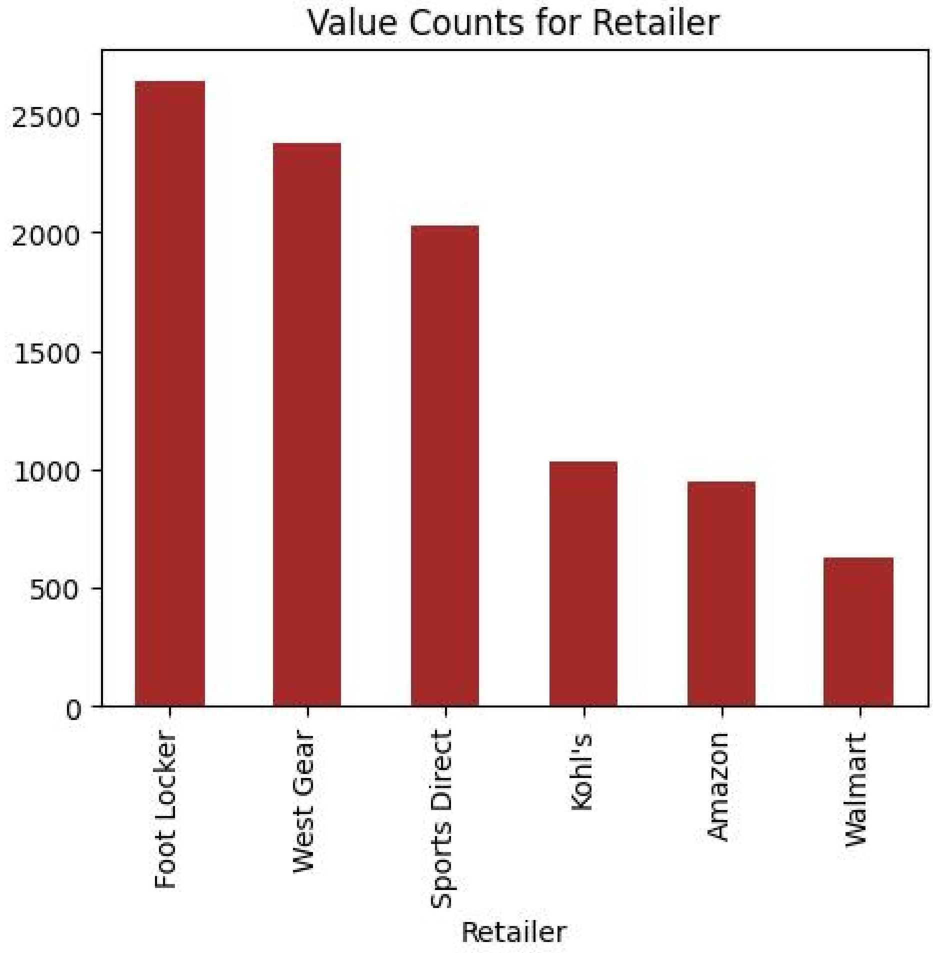

Figure 6.0 The bar chart shows the the count of sales for each retailer. This information can guide partnership decisions and incentive programs aimed at maximizing sales performance.

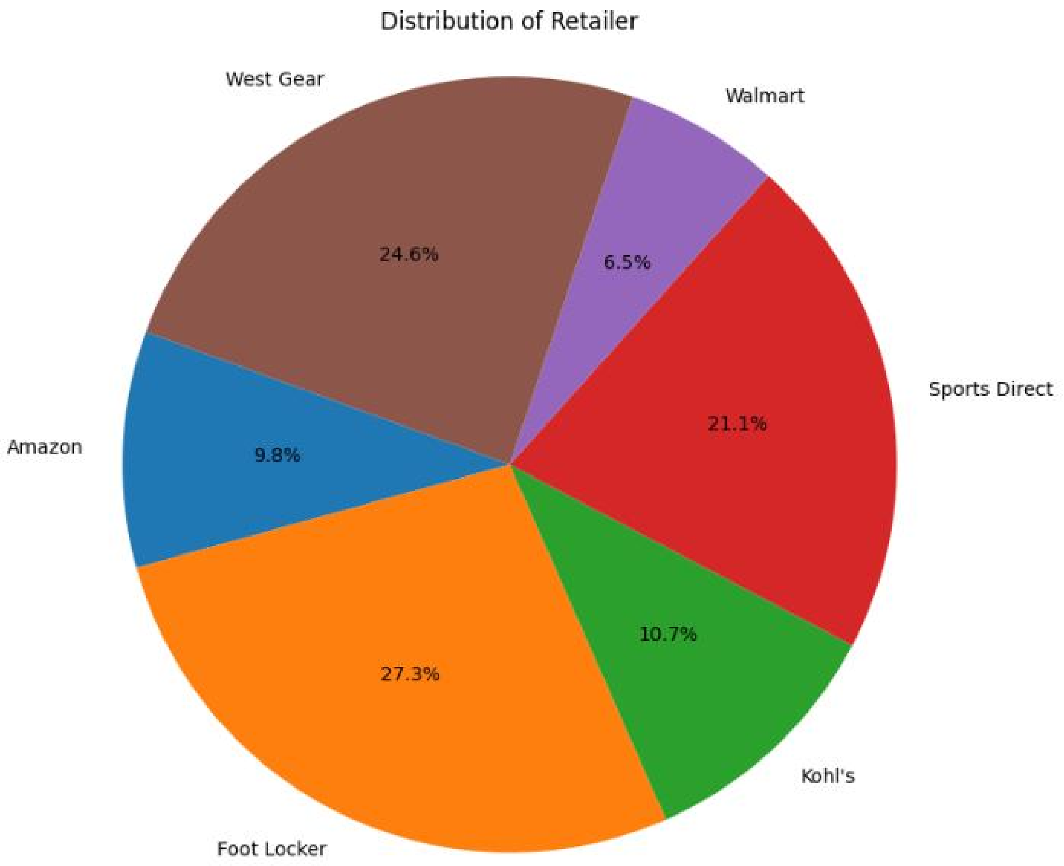

Figure 7.0 The pie chart shows the proportion of sales attributed to each retailer. This insight informs decisions related to retailer relationships, pricing strategies, and promotional activities.

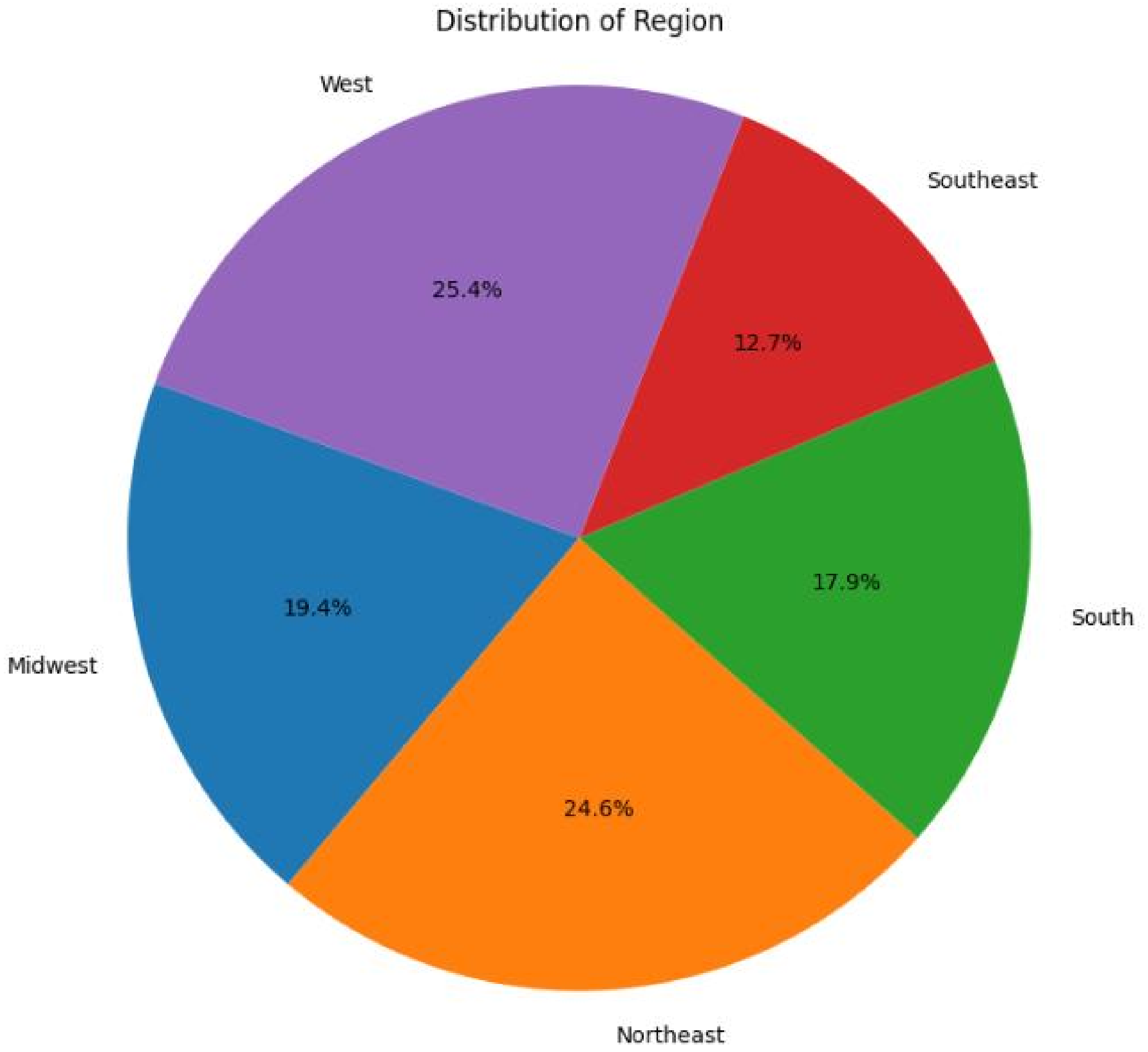

Figure 8.0 The pie chart the proportion of sales attributed to each region. This insight informs decisions related to regional expansion, inventory allocation, and market targeting.

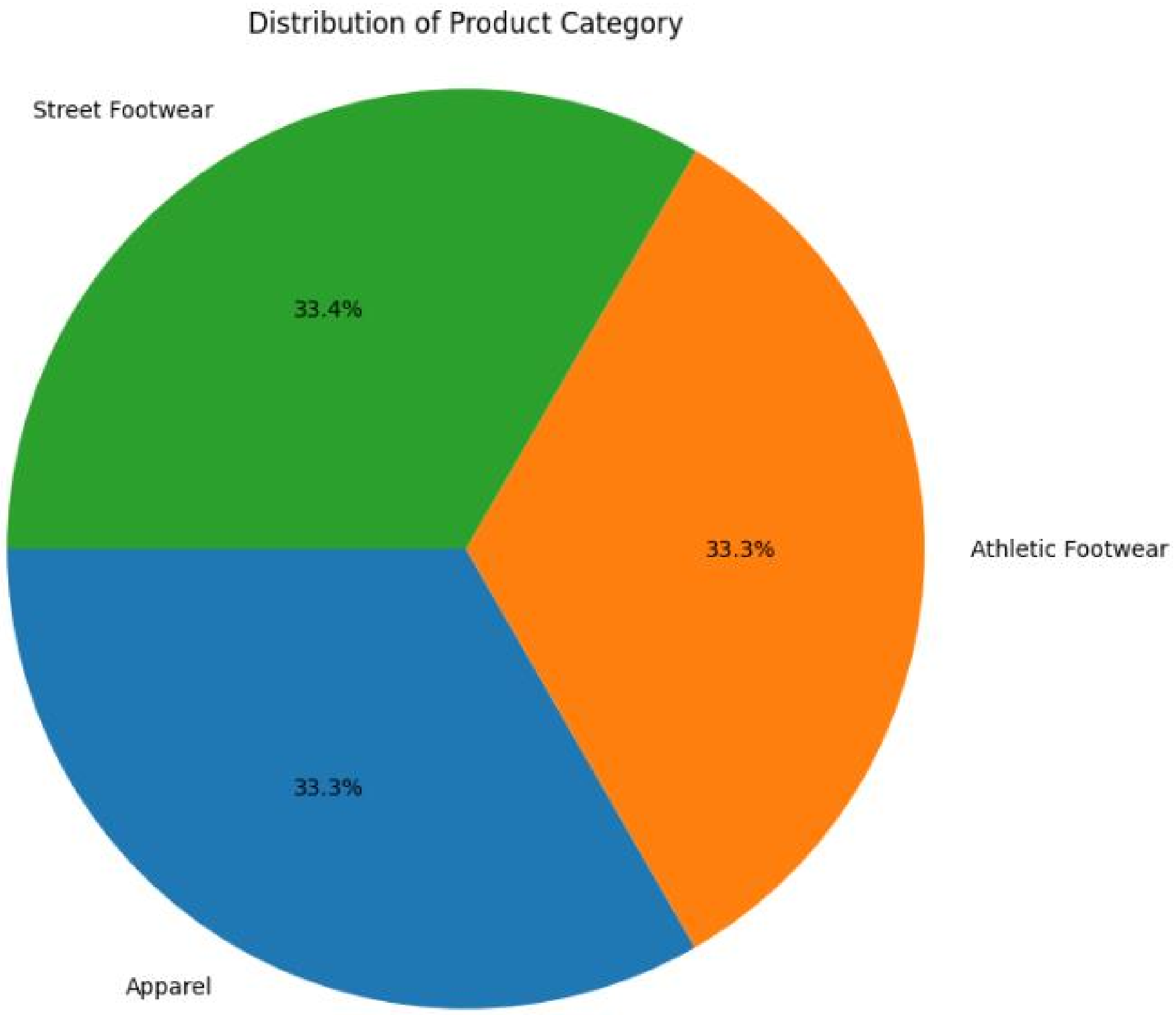

Figure 9.0 The pie chart shows the proportion of sales attributed to each product category. This insight helps in prioritizing marketing efforts, optimizing inventory management, and identifying opportunities for product diversification or expansion.

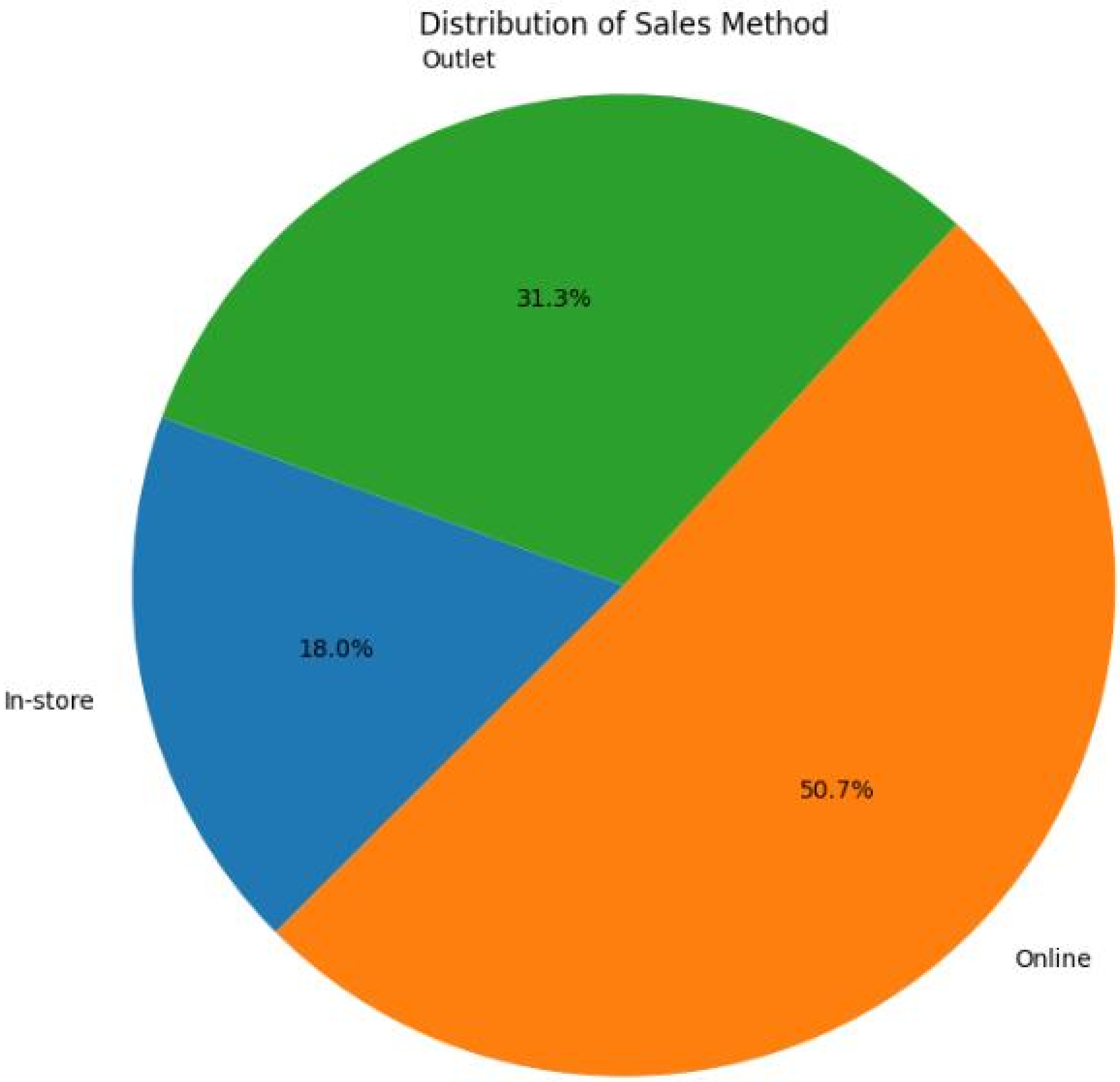

Figure 10.0 The pie chart shows the proportion of sales attributed to each sales method. This insight enables businesses to allocate resources strategically, focusing on the most profitable sales methods and optimizing their sales strategies accordingly.

Value Counts for Different Category:

- **Count Plots** – These plot visually represents the frequency of categorical data within the Adidas sales analysis dataset. Each bar corresponds to a category, with its height indicating the count or frequency of occurrences.

- **Implications** – Understanding the distribution patterns of categorical variables, aiding in identifying trends and making data-driven decisions. Also, will understand sales distribution across product categories guides inventory management and product development decisions.

Proportions for Different Category:

- **Pie chart** – visually represents the proportion of sales among different categories within the Adidas sales analysis dataset. Each slice of the pie corresponds to a category, with its size indicating the relative share of total sales attributed to that category.

- **Implications** – clear overview of the relative market share of each category, helping businesses understand the composition of their sales. For instance, visualizing the distribution of sales by product category can reveal which products are most popular, guiding marketing efforts and inventory planning.

VIII. Conclusion

This analysis of Adidas sales data offer significant value to the business by highlighting key trends and patterns across various dimensions such as region, city, retailer, product category, and sales method. By understanding the distribution and performance of sales in different geographic locations and through various sales channels, Adidas can tailor its marketing strategies, optimize inventory management, and enhance operational efficiency.

Data-driven decision-making is crucial for maintaining a competitive edge in the market. The ability to pinpoint high-performing regions and cities allows for targeted marketing and resource allocation, ensuring that efforts are concentrated where they will have the most impact. Also, understanding the popularity of different product categories aids in inventory planning and product development, reducing waste and meeting consumer demand more effectively.

Furthermore, the analysis of sales methods and retailer performance provides insights into the effectiveness of different sales channels and partnerships. This enables Adidas to refine its distribution strategies and develop more effective retailer relationships, ultimately driving higher sales and profitability.

The potential for future analysis is vast. Continual monitoring and analysis of sales data can uncover emerging trends, shifting consumer preferences, and new opportunities for growth. By integrating advanced machine learning techniques, Adidas can further enhance predictive capabilities, enabling proactive decision-making and strategic planning.