

# REGIONAL SALES ANALYSIS

## PROBLEM STATEMENT

- Sales teams often lack a clear, data-driven understanding of regional performance, making it difficult to identify growth opportunities and optimize resources. This project aims to analyse and visualize regional sales data to uncover trends, evaluate profitability, and support strategic decision-making.

# APPROACH

**Data Loading and Understanding**

**Data Cleaning and Preparation.**

**Exploratory Data Analysis (EDA)**

- **Univariate Analysis.**
- **Bivariate Analysis**

# DATA OVERVIEW

This project utilizes several datasets related to regional sales:

- Regional Sales Dataset (Sales Orders)**: Contains detailed information about individual sales orders, including order number, date, customer, channel, product, quantity, price, and cost.
- Regional Sales Dataset (Regions)**: Provides demographic and geographic information for various regions, such as population, median income, land area, and time zone.
- Regional Sales Dataset (Products)**: Lists product information, including product names.
- Regional Sales Dataset (Customers)**: Contains customer names.
- Regional Sales Dataset (State Regions)**: Maps state codes to their corresponding states and broader geographical regions (Midwest, Northeast, South, West).
- Regional Sales Dataset (2017 Budgets)**: Contains budget information for products in 2017.

These datasets were loaded, cleaned, and merged into a single comprehensive DataFrame (`df_merged`) to facilitate integrated analysis. The merged DataFrame contains information from all these sources, allowing for the exploration of data analysis phase.

# PROJECT WORKFLOW

Data Loading



Data cleaning  
and preparation



Data merging

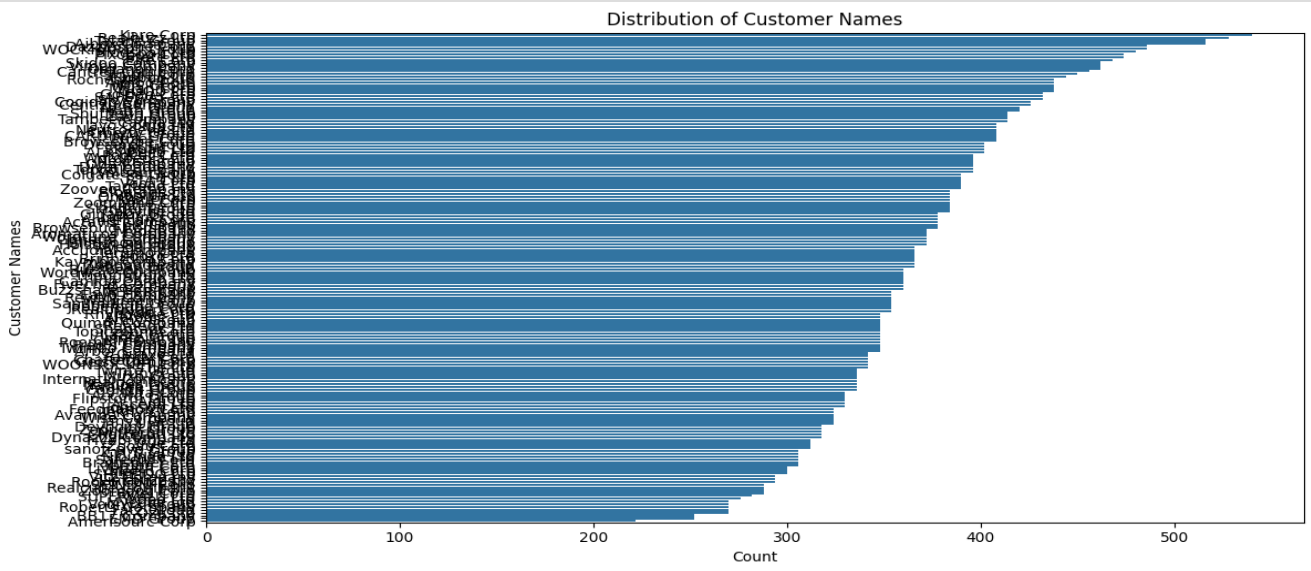
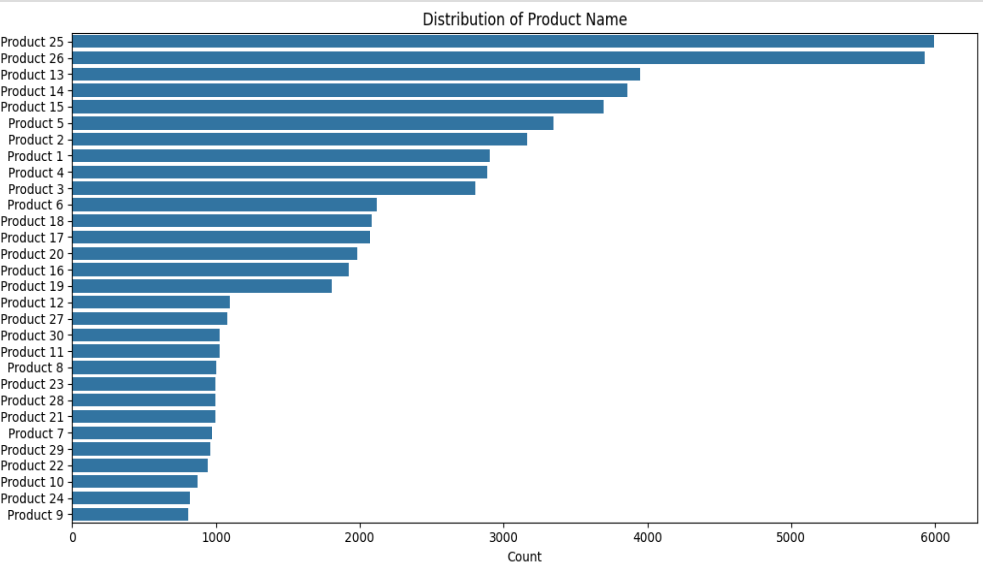
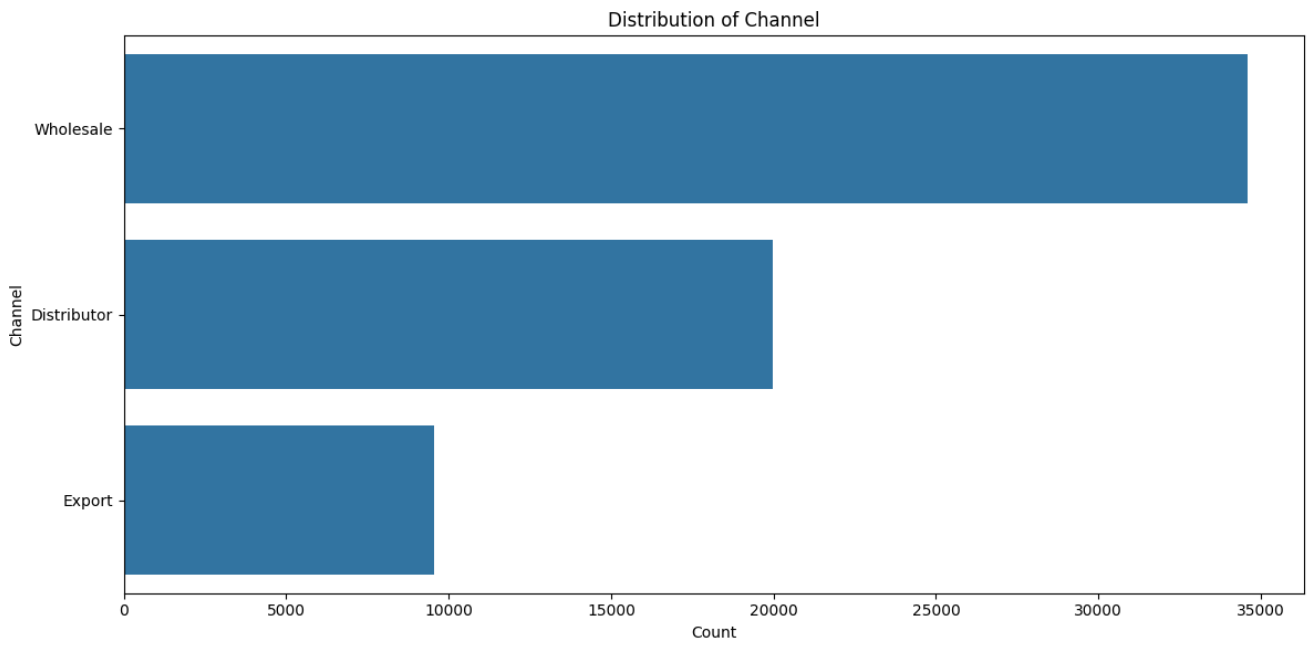
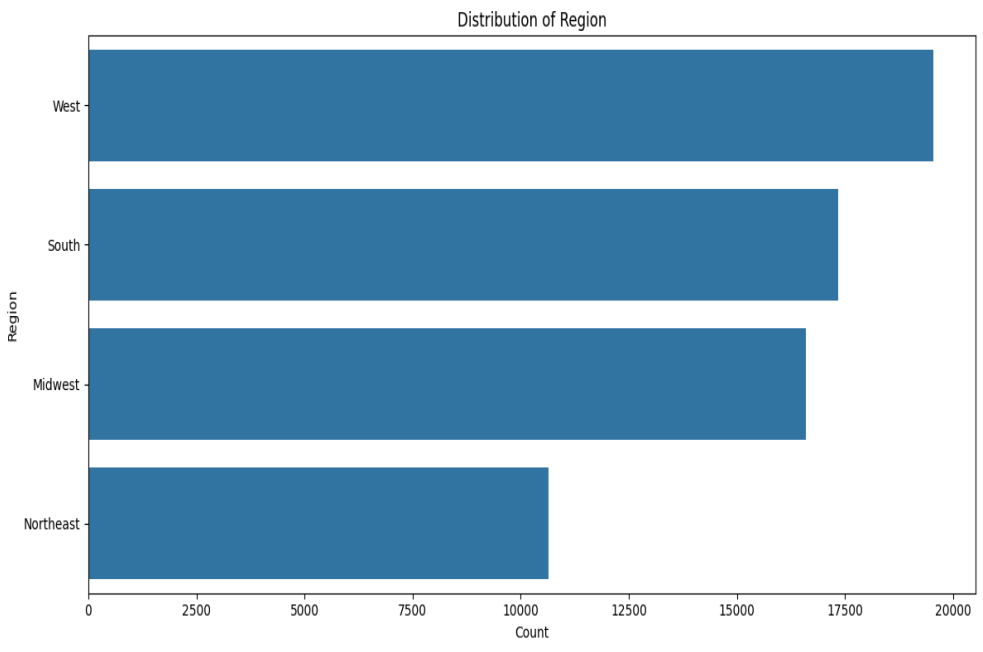


Exploratory  
data analysis



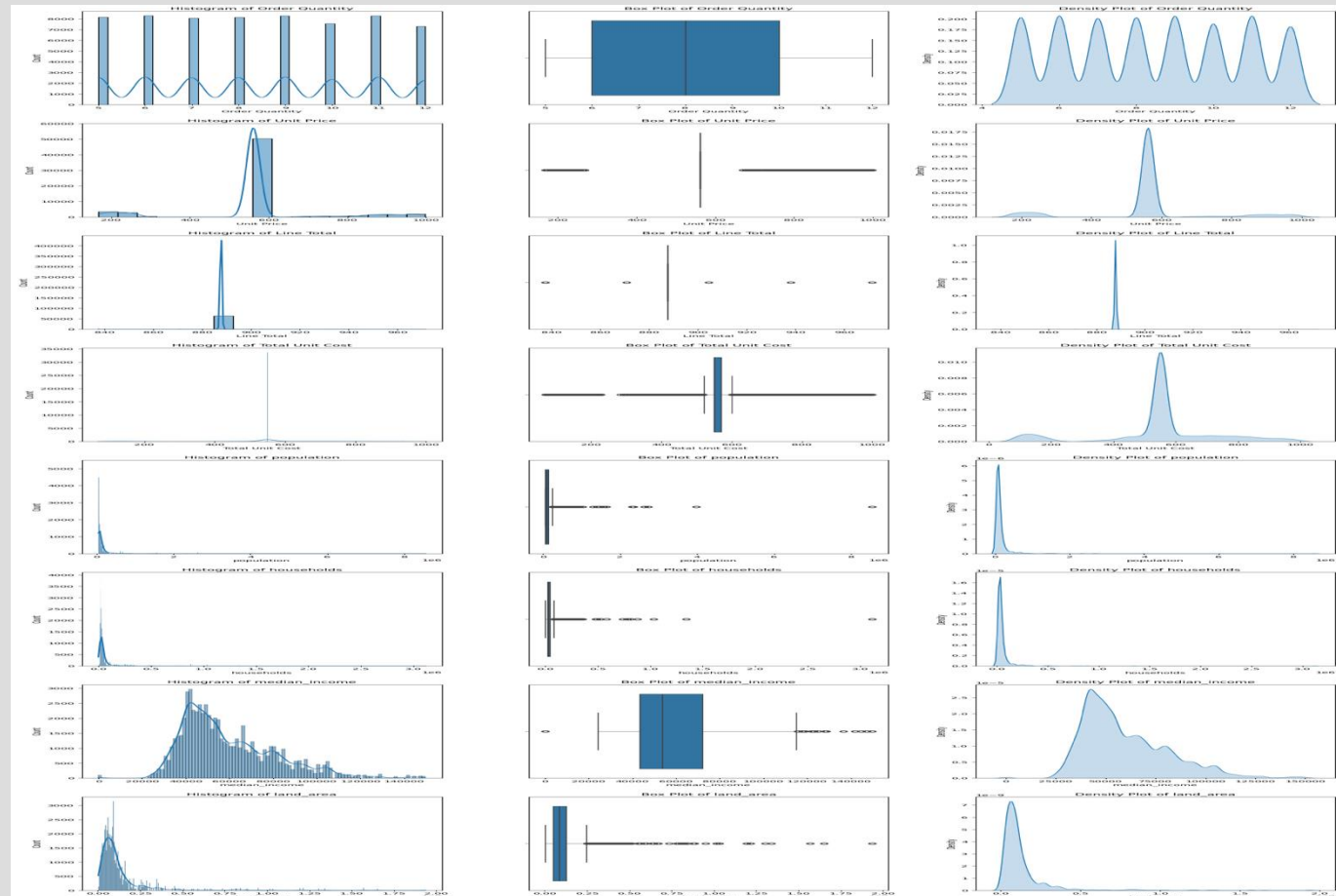
Insight  
generation

# Univariate analysis



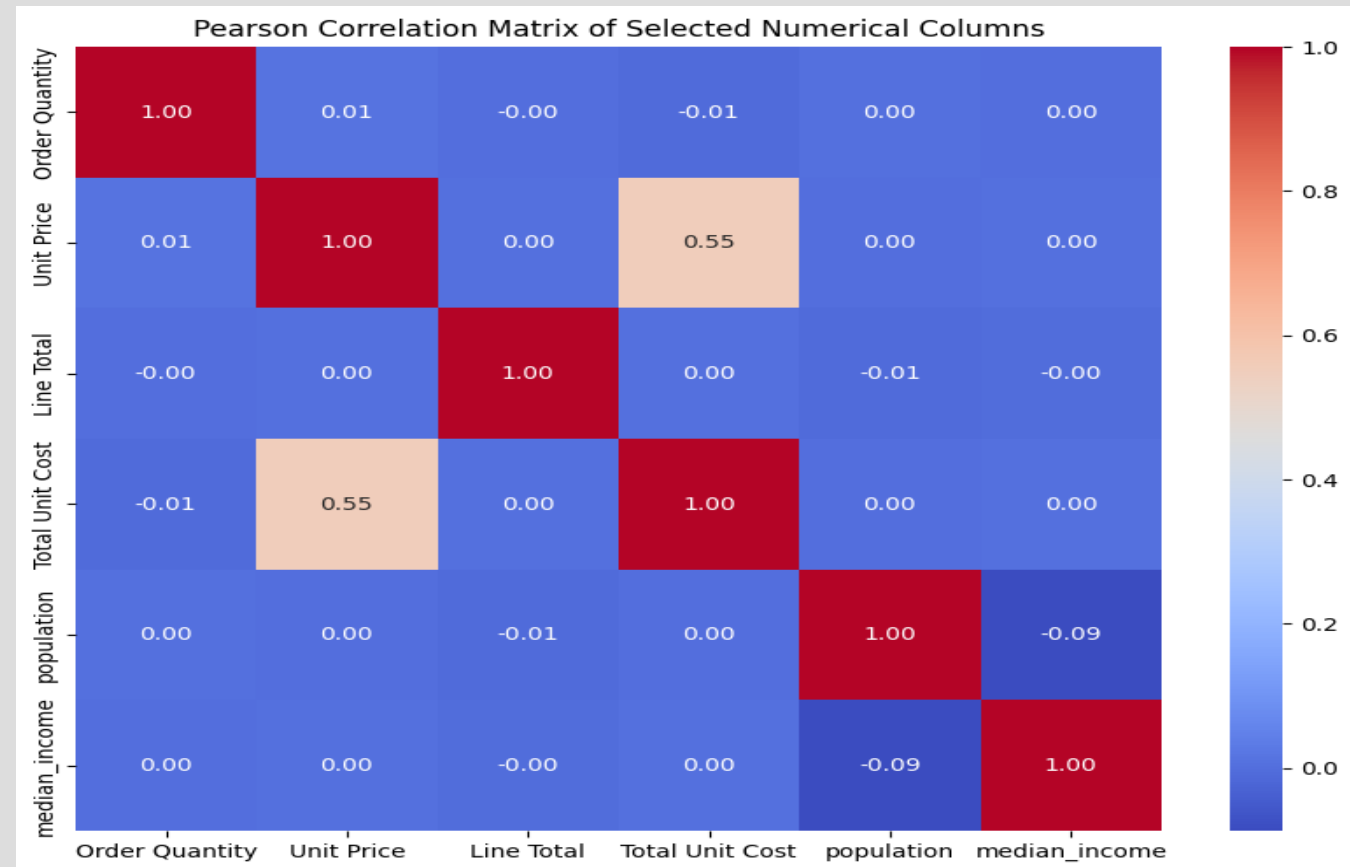
# HISTOGRAMS, BOX PLOTS, AND DENSITY PLOTS FOR A SELECTION OF KEY NUMERICAL COLUMNS TO VISUALIZE THEIR DISTRIBUTIONS AND IDENTIFY POTENTIAL OUTLIERS

- The numerical analysis provides a good understanding of the central tendencies, spread, and potential presence of outliers in these key quantitative aspects of the data. This information is valuable when considering how these factors might influence sales or other aspects of the dataset.



# RELATIONSHIPS BETWEEN NUMERICAL COLUMNS

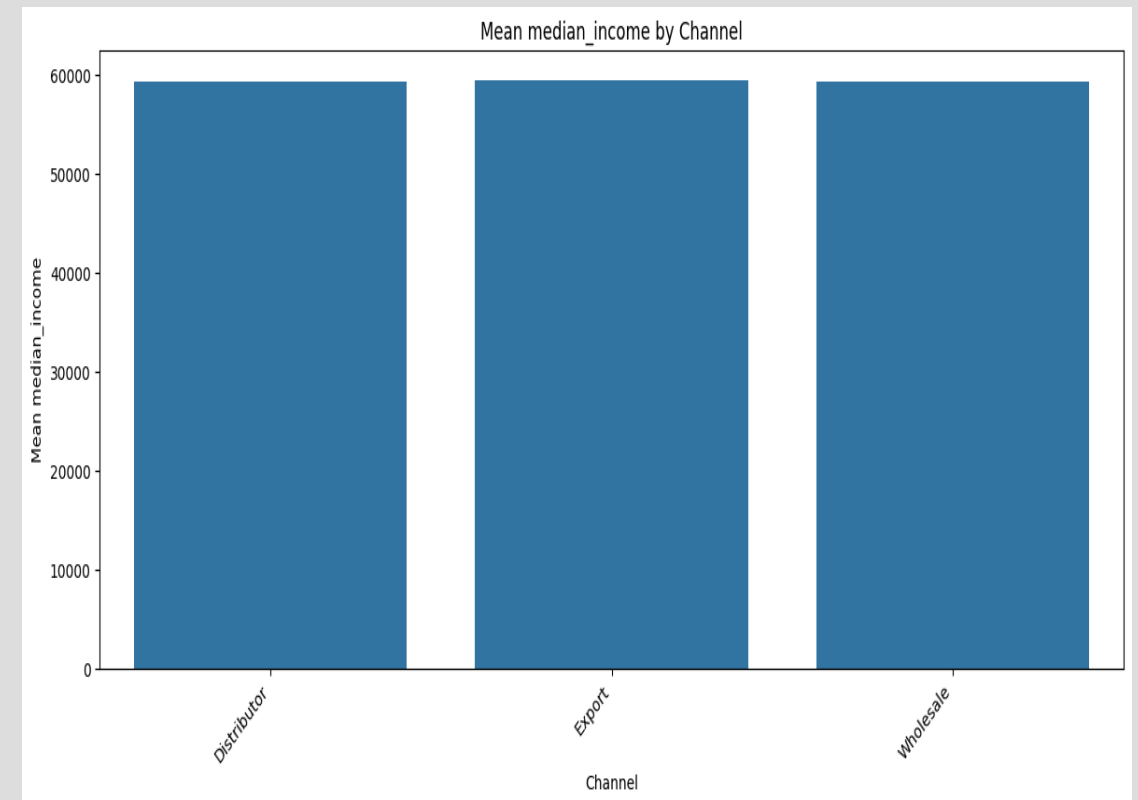
- The Pearson correlation matrix for the selected numerical columns and visualize it using a heatmap to understand linear relationships.
- The heatmap primarily highlights the moderate positive relationship between 'Unit Price' and 'Total Unit Cost'. For most other selected numerical pairs, there isn't a strong indication of a linear relationship based on the correlation coefficients.





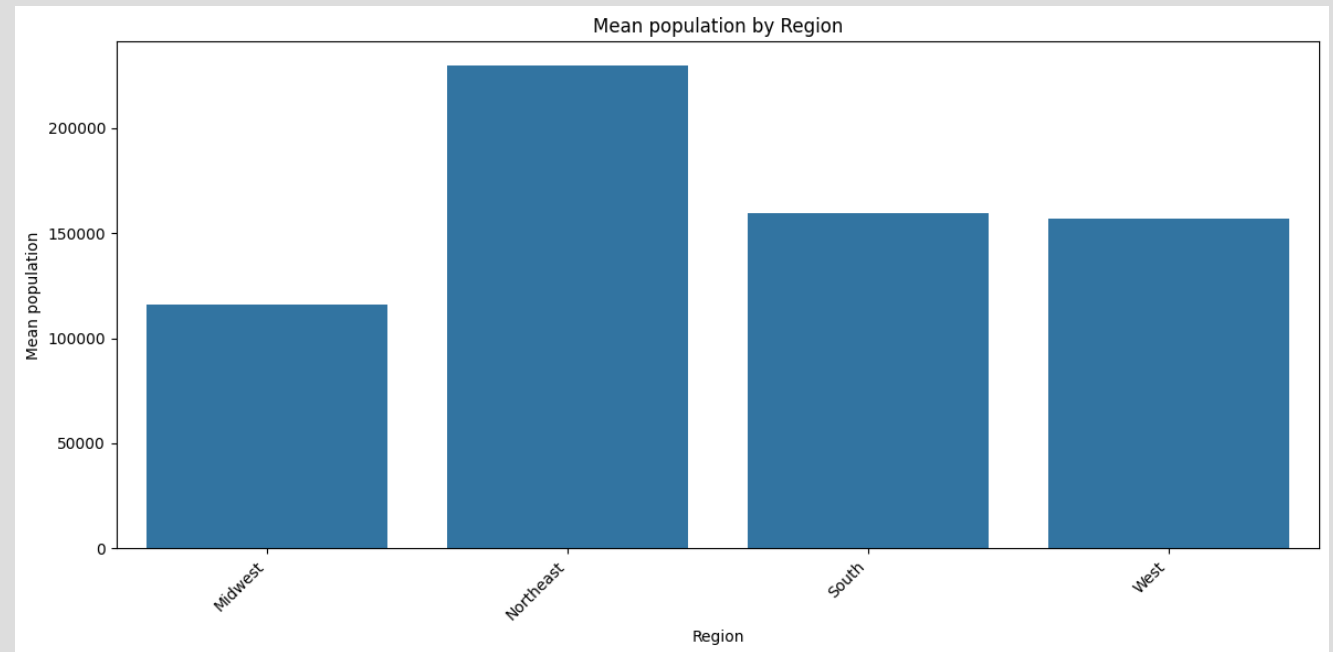
## BAR PLOT SHOWING THE MEAN MEDIAN INCOME BY CHANNEL

This visualization provides a clear and concise summary of the economic profile of the areas where each channel operates, based on the mean median income. It complements the box plot by showing the average rather than the distribution, giving you a different perspective on the relationship between channel and regional income levels.



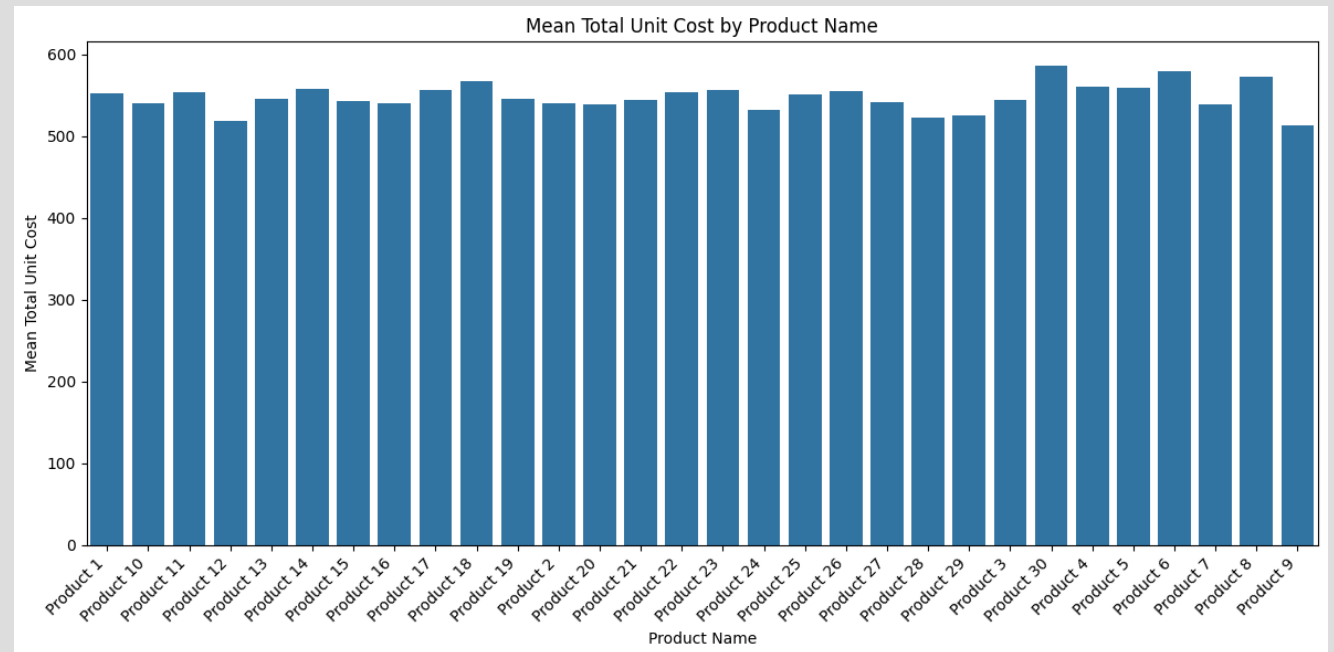
## BAR PLOT SHOWING THE MEAN POPULATION BY REGION

This visualization provides a clear summary of the average population density of the areas where you have sales in each geographical region. It complements the box plot by showing the average rather than the distribution, giving you a different perspective on the relationship between region and population levels.



## BAR PLOT SHOWING THE MEAN TOTAL UNIT COST BY PRODUCT NAME

This visualization provides a clear way to compare the average cost of goods for each product in your dataset. This information is valuable for understanding the cost structure of your products and can be used in conjunction with pricing and sales data to analyze profitability by product.

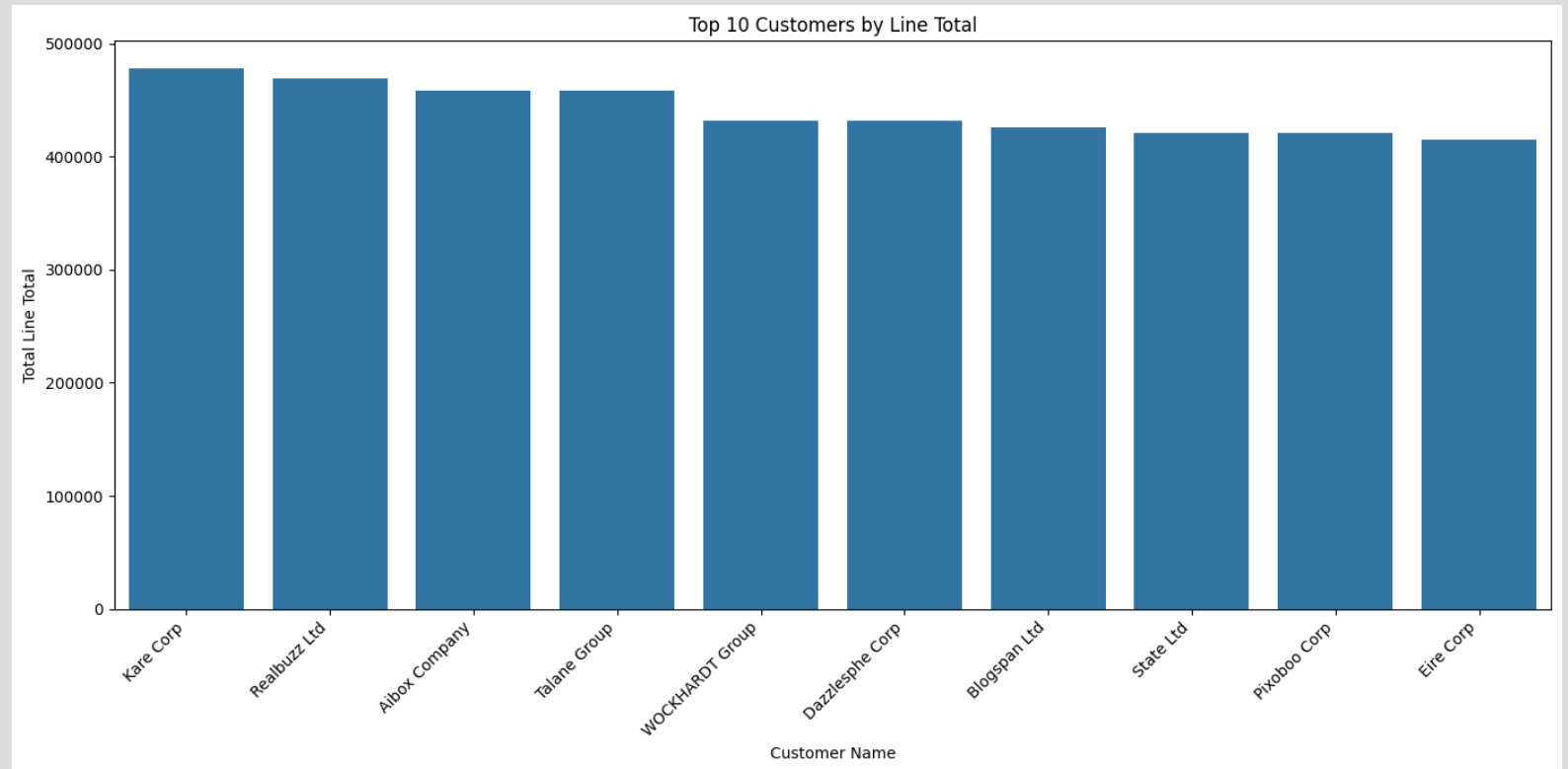


# TOP 10 CUSTOMERS

The top 10 customers by Line Total are:

- Kare Corp: 478480.5
- Realbuzz Ltd: 469033.5
- Aibox Company: 458079.0
- Talane Group: 458079.0
- WOCKHARDT Group: 432150.0
- Dazzlesphe Corp: 431647.5
- Blogspan Ltd: 426120.0
- State Ltd: 420793.5
- Pixoboo Corp: 420793.5
- Eire Corp: 415467.0

These customers contribute the most to the overall Line Total. Focusing on retaining and potentially expanding business with these customers could be a key strategy for increasing revenue. Further analysis could involve understanding the characteristics of these top customers, such as their location, the products they purchase most frequently, and their preferred channels, to tailor marketing and sales efforts.

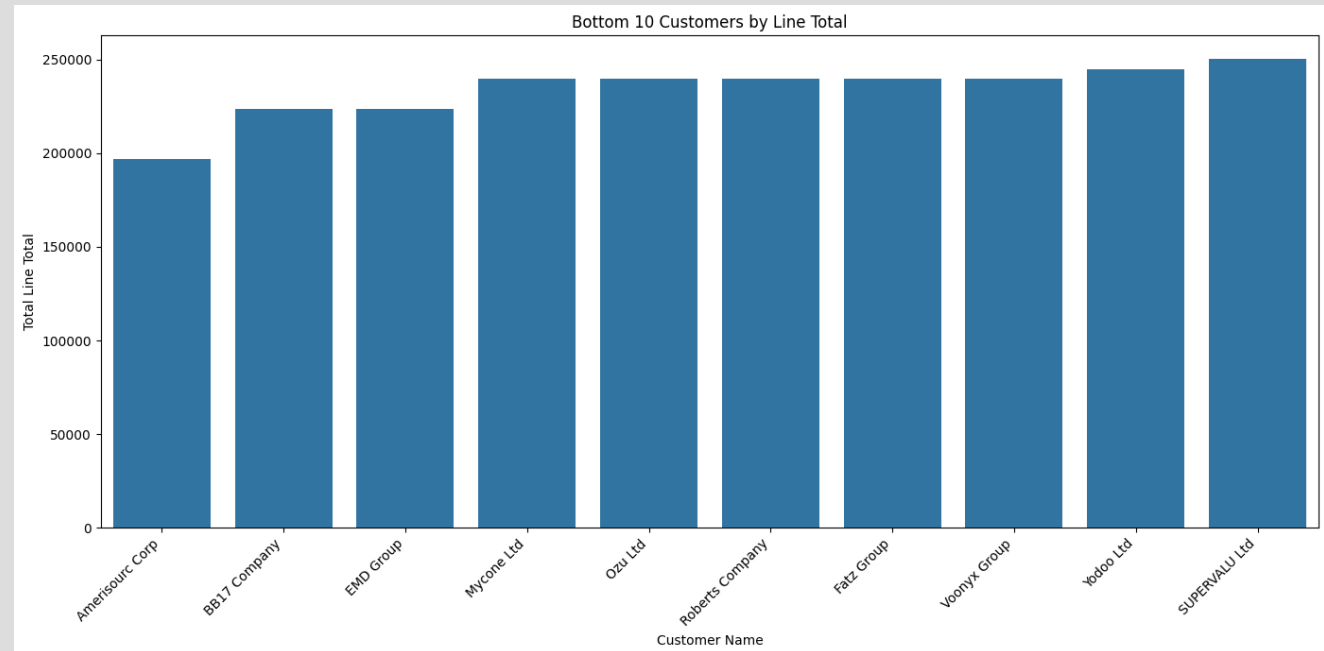


# BOTTOM 10 CUSTOMERS

The bottom 10 customers by Line Total are:

- Amerisourc Corp: 197080.5
- BB17 Company: 223713.0
- EMD Group: 223713.0
- Mycone Ltd: 239592.0
- Ozu Ltd: 239692.5
- Roberts Company: 239692.5
- Fatz Group: 239692.5
- Voonyx Group: 239692.5
- Yodoo Ltd: 245019.0
- SUPERVALU Ltd: 250345.5

These customers contribute the least to the overall Line Total. Understanding the reasons behind their low Line Total could be beneficial. It could be due to various factors such as their size, purchase frequency, product preferences, or other issues. Further investigation into these customers might reveal opportunities for improvement or indicate that they are not a primary focus for growth.



# TOP 10 STATES

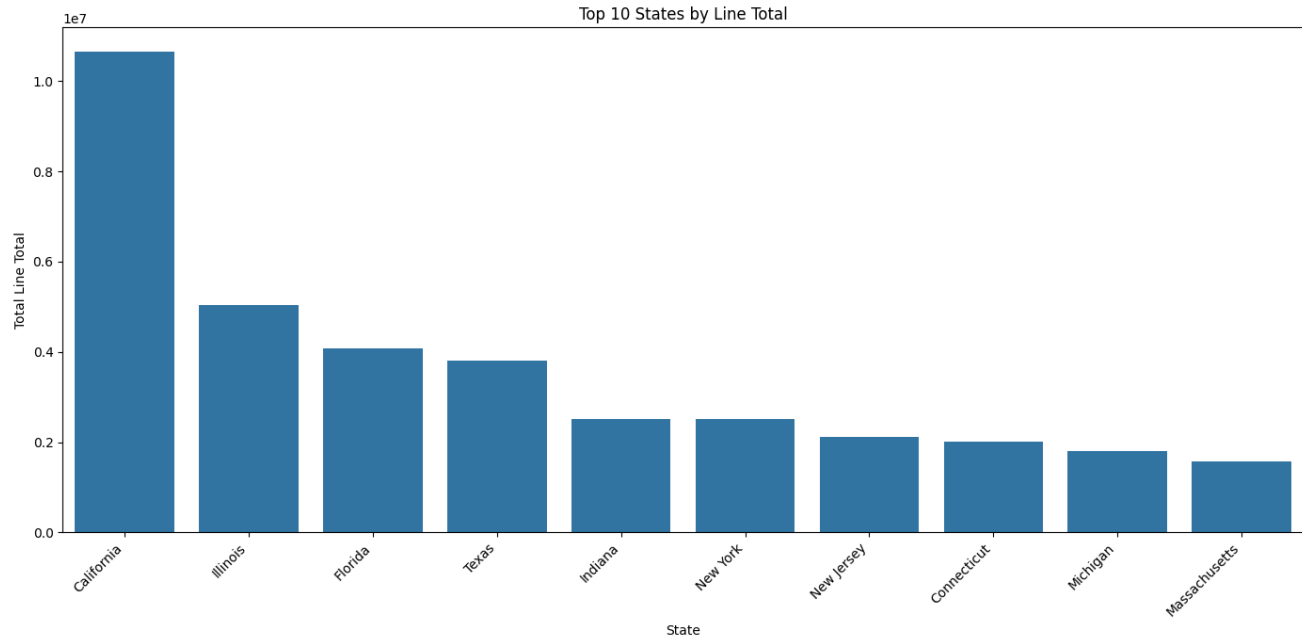
- California is the clear leader**
- Illinois and Florida are strong contenders**

- Texas is also a major contributor:** Texas is the fourth state with a substantial line total, reinforcing its importance in sales.

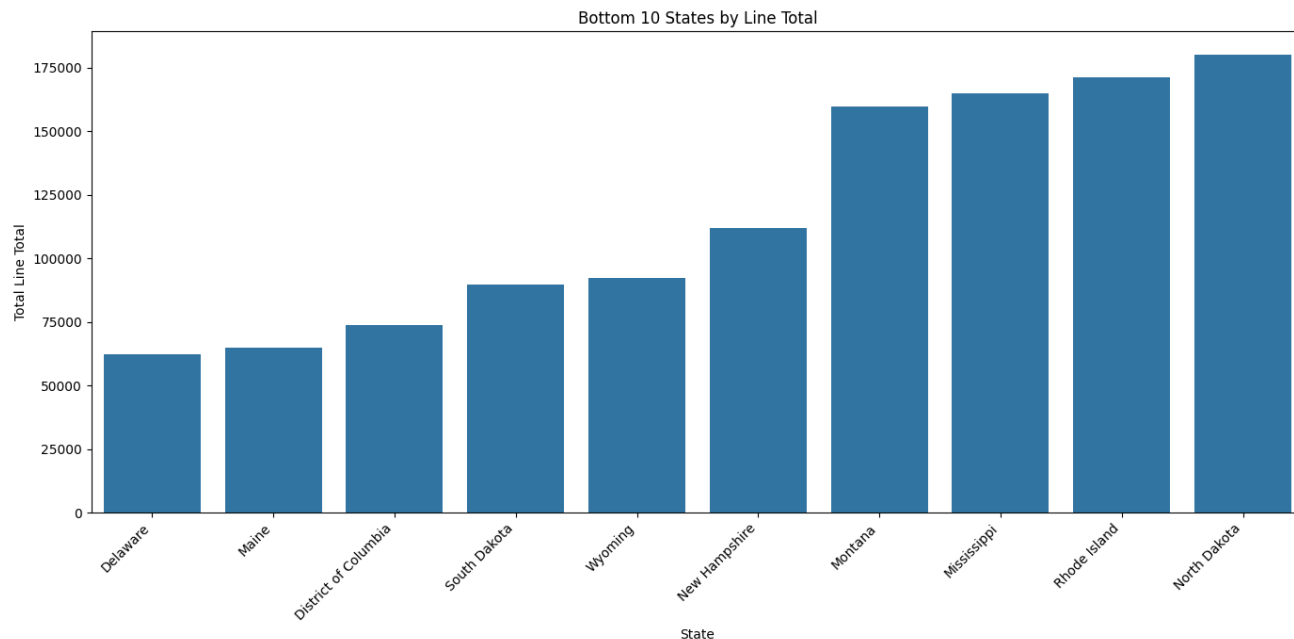
- Concentration of sales:** The top few states contribute a large portion of the total line total, suggesting a concentration of sales in specific geographical areas.

- Gradual decrease:** After the top few states, there is a more gradual decrease in the total line total among the remaining states in the top 10.

These insights suggest that California is a key market, and focusing on strategies to maintain or increase sales in California, Illinois, Florida, and Texas could be beneficial. Further analysis could explore why these states have such high line totals and if there are specific products or channels that perform particularly well in these regions.



## BOTTOM 10 STATES



- Very low line totals:** The states in the bottom 10 have significantly lower line totals compared to the top states, highlighting a large disparity in sales performance across different states.

- Delaware and Maine are the lowest:** Delaware and Maine have the lowest total line totals among all states in the dataset.

- Similar performance among many states:** Several states in the bottom 10 have relatively similar low line totals, indicating a cluster of states with minimal sales contribution.

- Potential areas for growth:** These states represent potential areas where sales strategies could be implemented or improved to increase revenue.

These insights suggest that there are states with very low sales activity. Understanding the reasons behind these low numbers (e.g., market size, competition, product demand, sales channel effectiveness) could inform strategies for improvement.

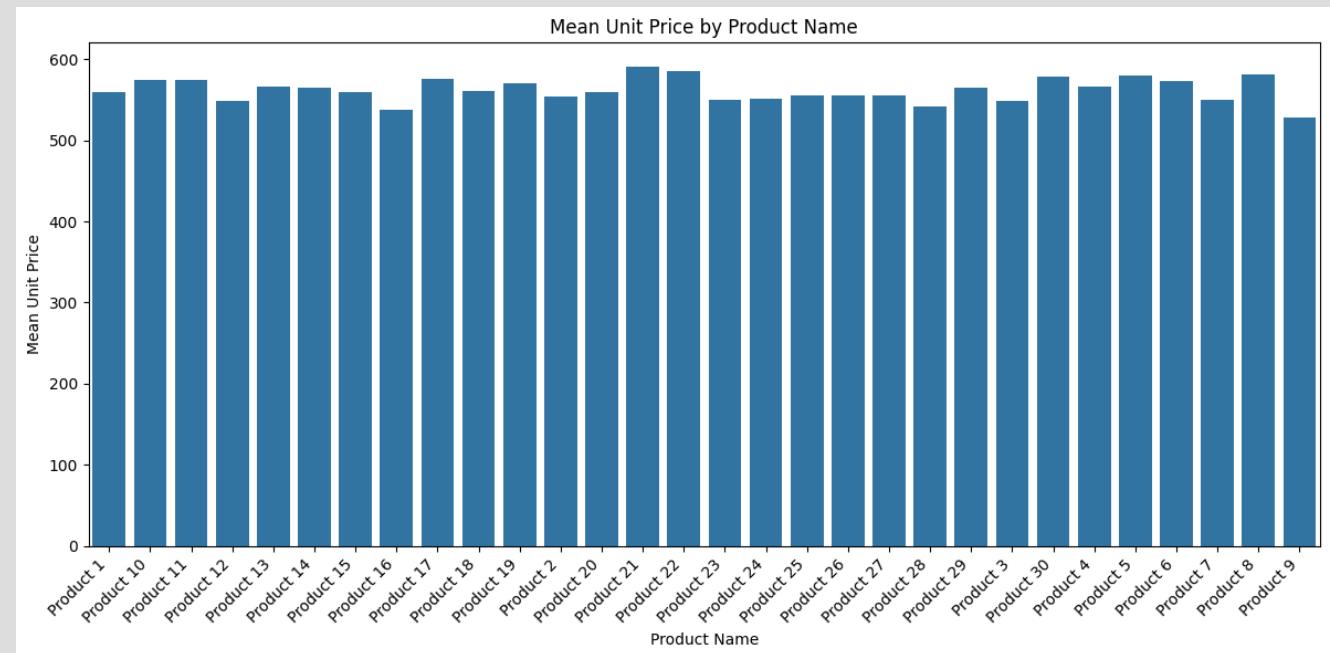
# UNIT PRICE BY PRODUCT DISTRIBUTION

- Variability in Unit Prices:** There is a notable variation in the average unit prices across different products. Some products have significantly higher mean unit prices than others.

- Highest Average Unit Prices:** Products like 'Product 11', 'Product 17', 'Product 21', 'Product 22', 'Product 5', and 'Product 8' appear to have higher average unit prices compared to the rest.

- Lowest Average Unit Prices:** Products like 'Product 16', 'Product 9', and 'Product 23' seem to have lower average unit prices.

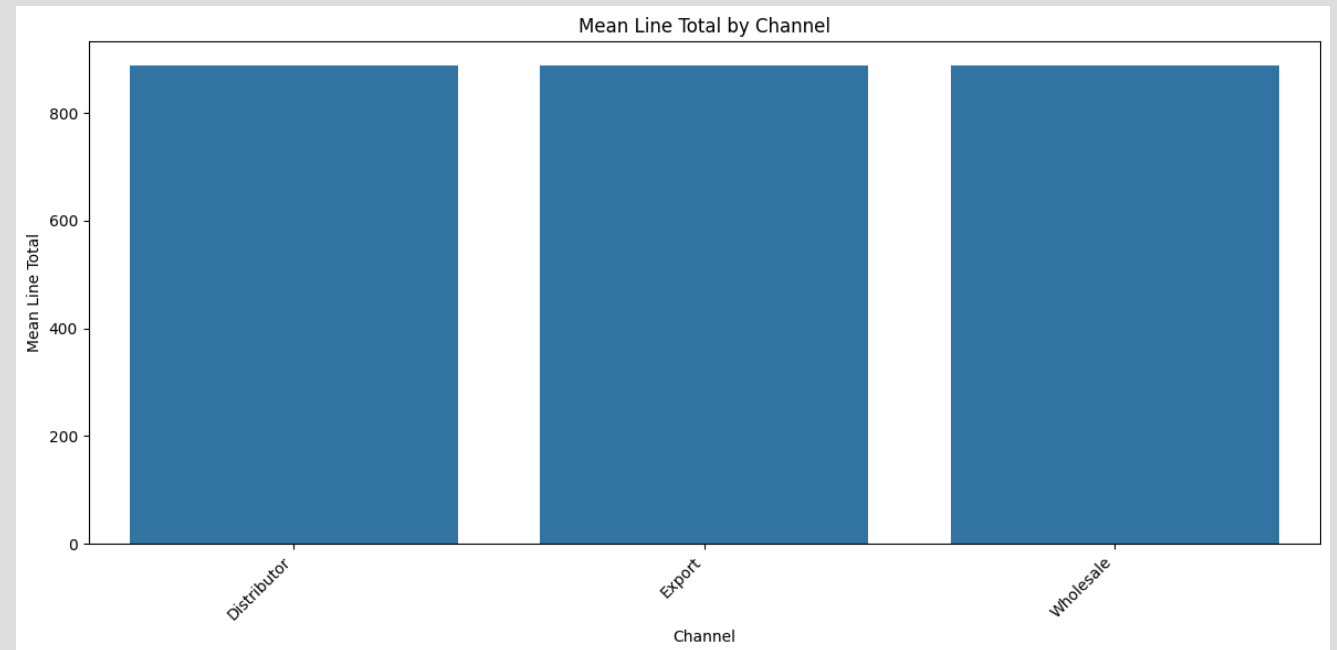
- Most Products Clustered Around the Mean:** A large number of products have mean unit prices that are relatively close to the overall average unit price, indicating a central tendency in pricing for many items.. These insights highlight that the unit price is not uniform across all products. Understanding which products command higher or lower average unit prices can be valuable for pricing strategies, product development, and sales forecasting.





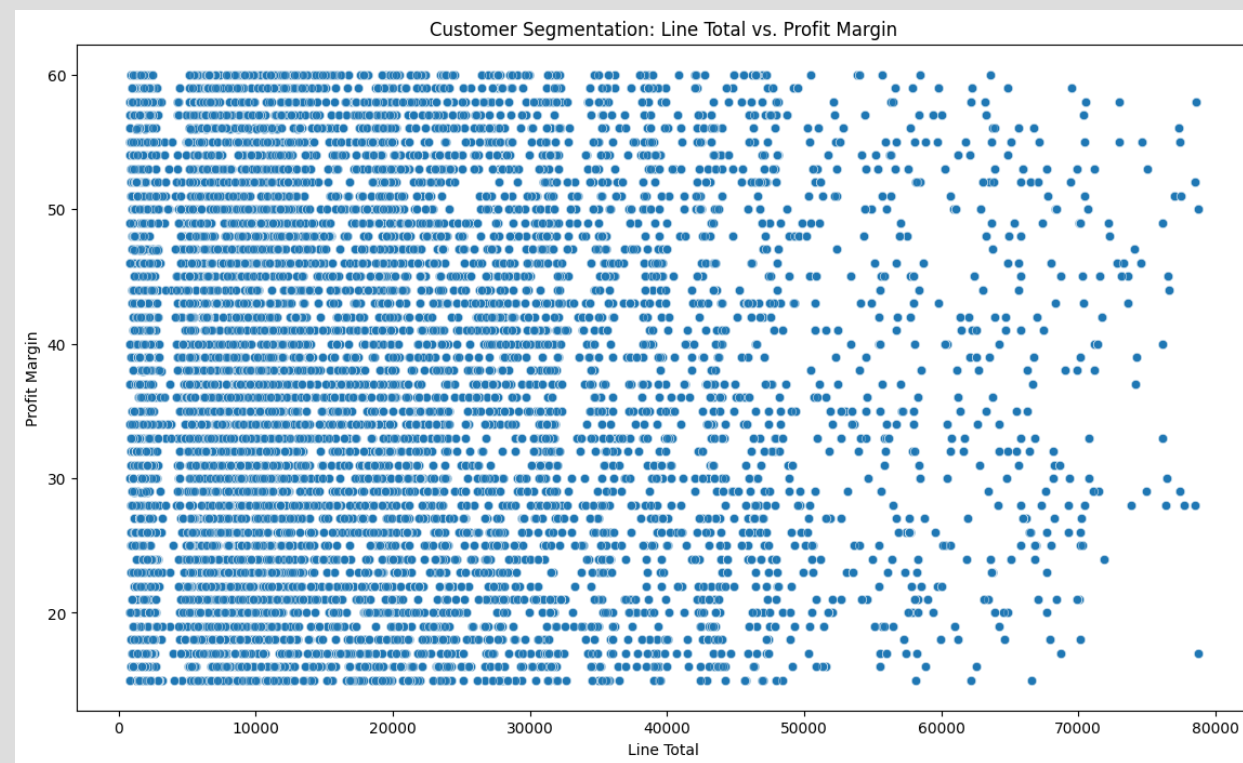
# LINE TOTAL BY CHANNEL DISTRIBUTION

- Similar Average Line Totals:** The average line totals across the 'Distributor', 'Export', and 'Wholesale' channels appear to be quite similar. There isn't a significant difference in the average revenue generated per order among these channels.
  - No Single Channel Dominance (on average):** Unlike some other metrics where one channel might clearly outperform others, the mean line total suggests that on average, each channel contributes a comparable amount of revenue per transaction.
  - Implication for Channel Strategy:** Since the average order value is similar, the overall revenue generated by each channel will be primarily driven by the *volume* of orders rather than the size of individual orders.
  - Focus on Volume:** To increase overall revenue, the focus for all channels might be on increasing the number of orders rather than trying to significantly increase the value of each individual order, based on this specific metric.
- These insights suggest that while the channels may differ in other aspects (like total sales volume or types of customers), the average value of a transaction is relatively consistent across 'Distributor', 'Export', and 'Wholesale'.



# LINE TOTAL VS PROFIT MARGIN

- Each point is a customer:** Every dot on the plot represents a single customer's overall performance based on these two metrics.
- X-axis (Line Total):** As you move to the right on the x-axis, the customer's total revenue from orders increases.
- Y-axis (Profit Margin):** As you move up on the y-axis, the customer's profit margin increases (meaning they are more profitable). Negative values on the y-axis indicate that the total cost of their orders was higher than the revenue generated.
- Identifying Segments:** You can look for clusters or patterns of points to identify different customer segments:
  - High Revenue, High Profit:** Points in the upper right corner represent customers who generate high revenue and are highly profitable. These are likely your most valuable customers.
  - High Revenue, Low/Negative Profit:** Points in the lower right corner indicate customers who generate high revenue but are not very profitable, or even unprofitable. Understanding the reasons for low profit margin for these high-revenue customers is crucial.
  - Low Revenue, High Profit:** Points in the upper left corner represent customers who don't generate a lot of revenue but are highly profitable on the business they do.
  - Low Revenue, Low/Negative Profit:** Points in the lower left corner represent customers who generate low revenue and are not profitable.
- Outliers:** Individual points that are far away from the main clusters might represent unique customer behaviors or potential data anomalies.



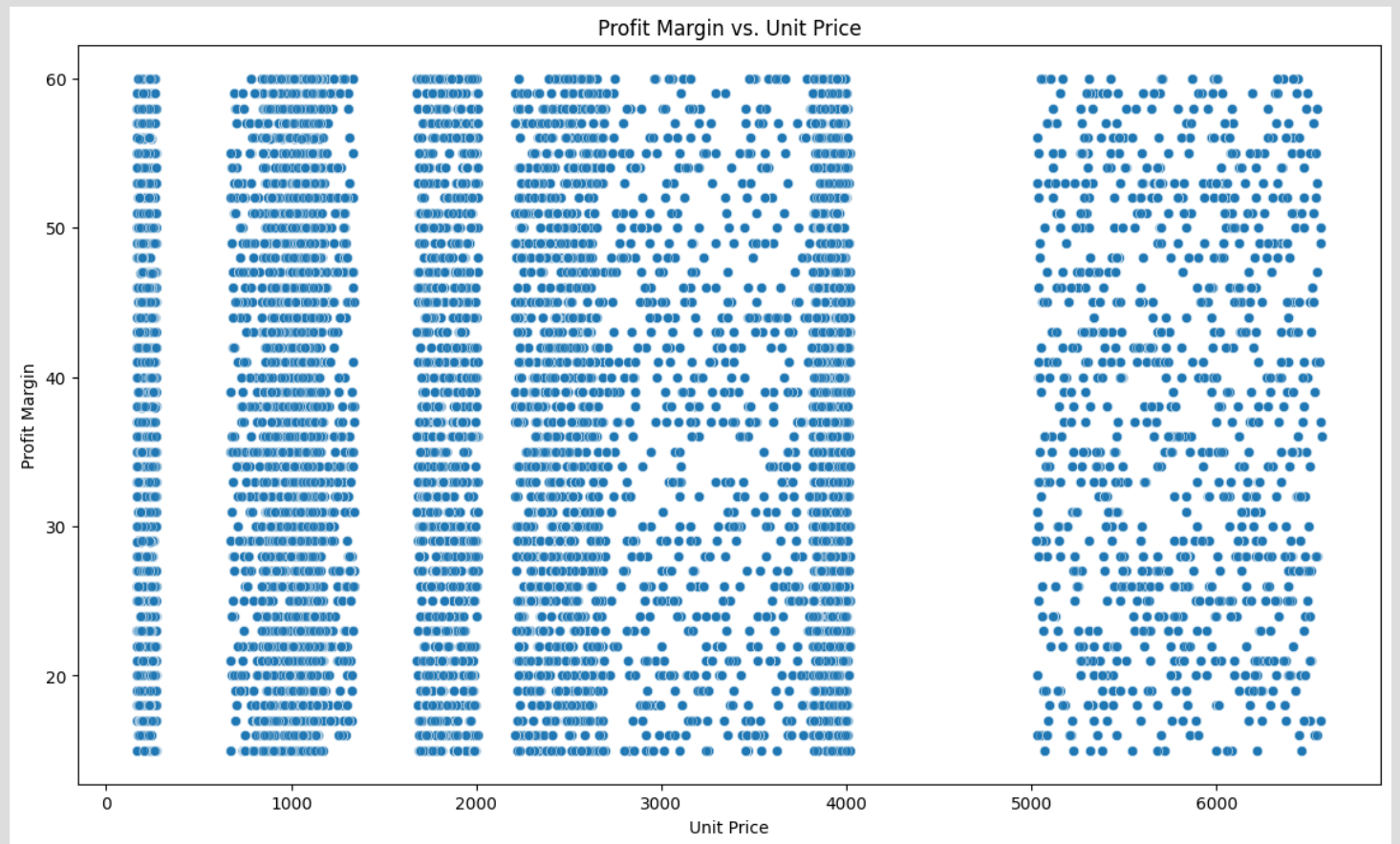
# PROFIT MARGIN VS UNIT PRICE

- No strong linear relationship:** There doesn't appear to be a strong positive or negative linear correlation between Unit Price and Profit Margin. The points are scattered, suggesting that a higher unit price doesn't automatically guarantee a higher profit margin, nor does a lower unit price necessarily mean a lower profit margin.

- Variability in Profit Margin:** For any given unit price, there is a wide range of profit margins. This indicates that other factors besides the unit price significantly influence the profit margin.

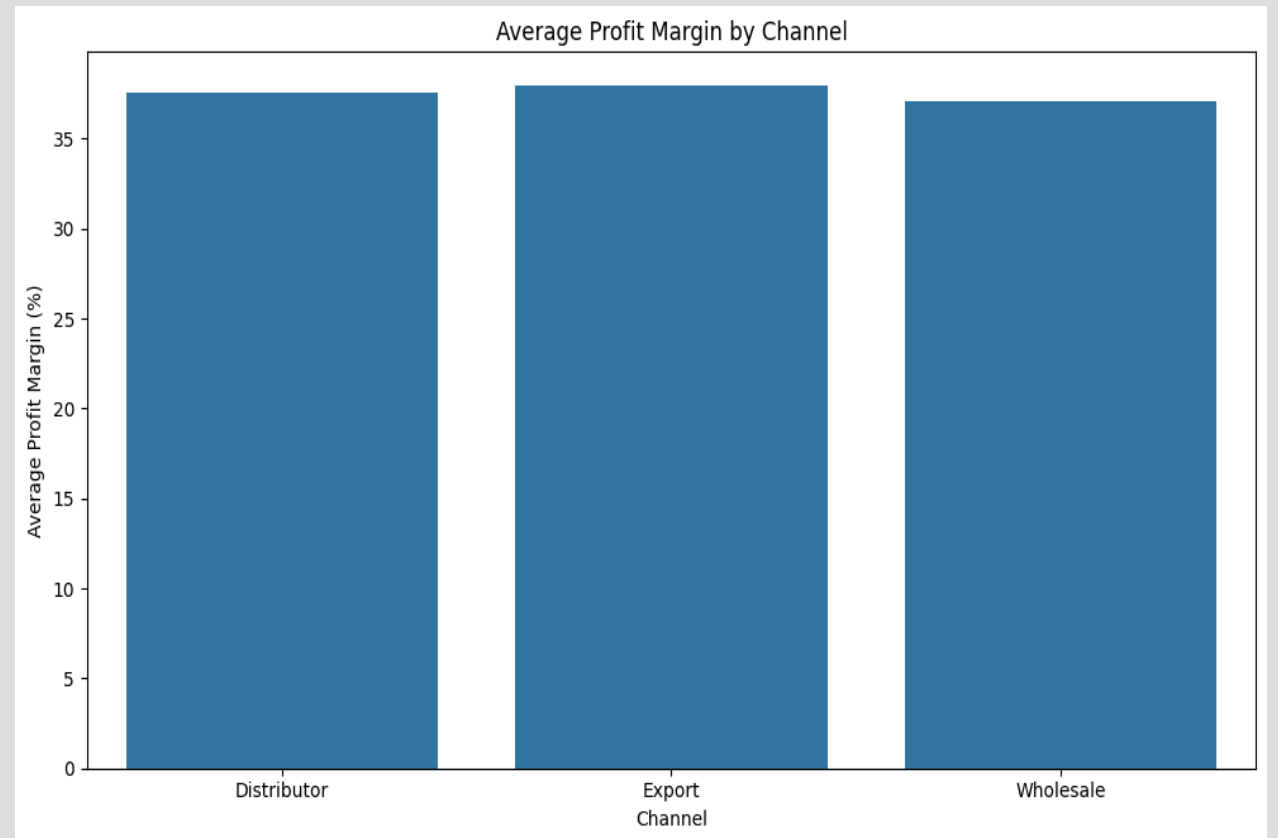
- Potential clusters or patterns:** While there isn't a clear overall trend, there might be some clusters or groups of data points. For example, some products with lower unit prices might have a wide spread of profit margins, while products with very high unit prices might have a slightly more constrained range of profit margins. Further investigation into these clusters could reveal interesting patterns related to specific product types, channels, or regions.

- Outliers:** There may be some outliers where a product has a very high or very low profit margin relative to its unit price. These outliers could be worth investigating to understand the reasons behind their unusual profit performance.



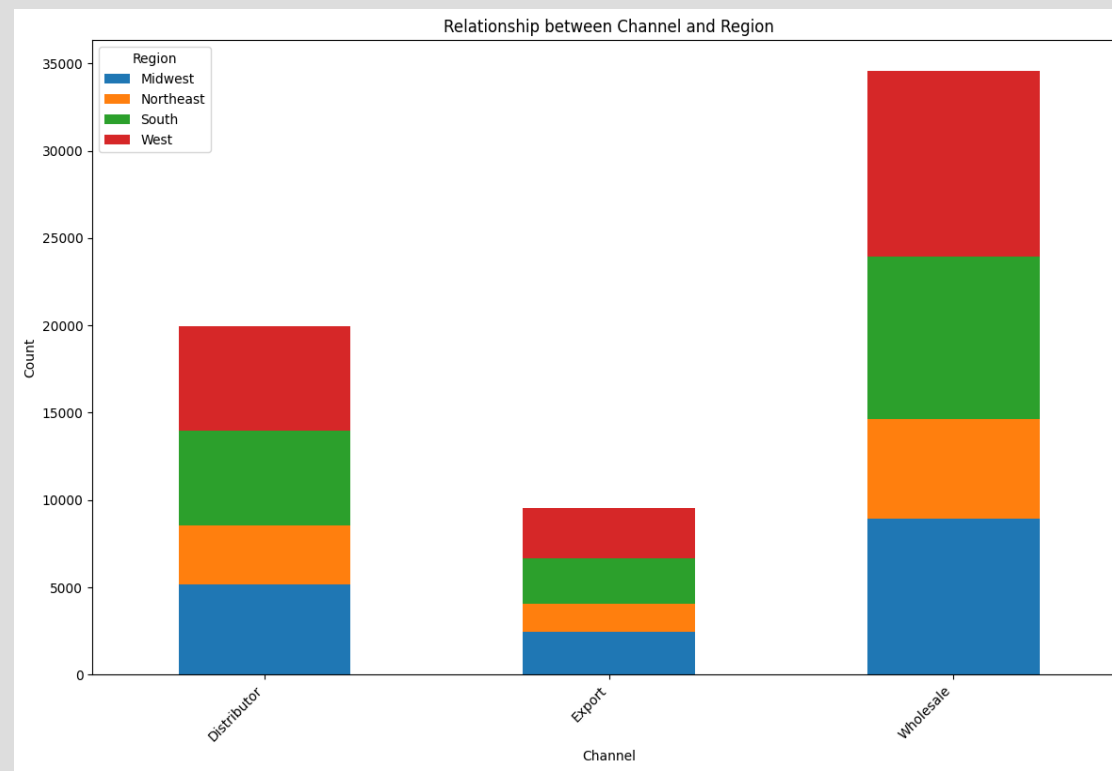
## AVERAGE PROFIT MARGIN VS CHANNEL

While the differences in average profit margin between channels are not very large, these variations could still be significant depending on the overall sales volume of each channel. To get a more complete picture, we could also analyze the total profit generated by each channel, not just the average profit margin.



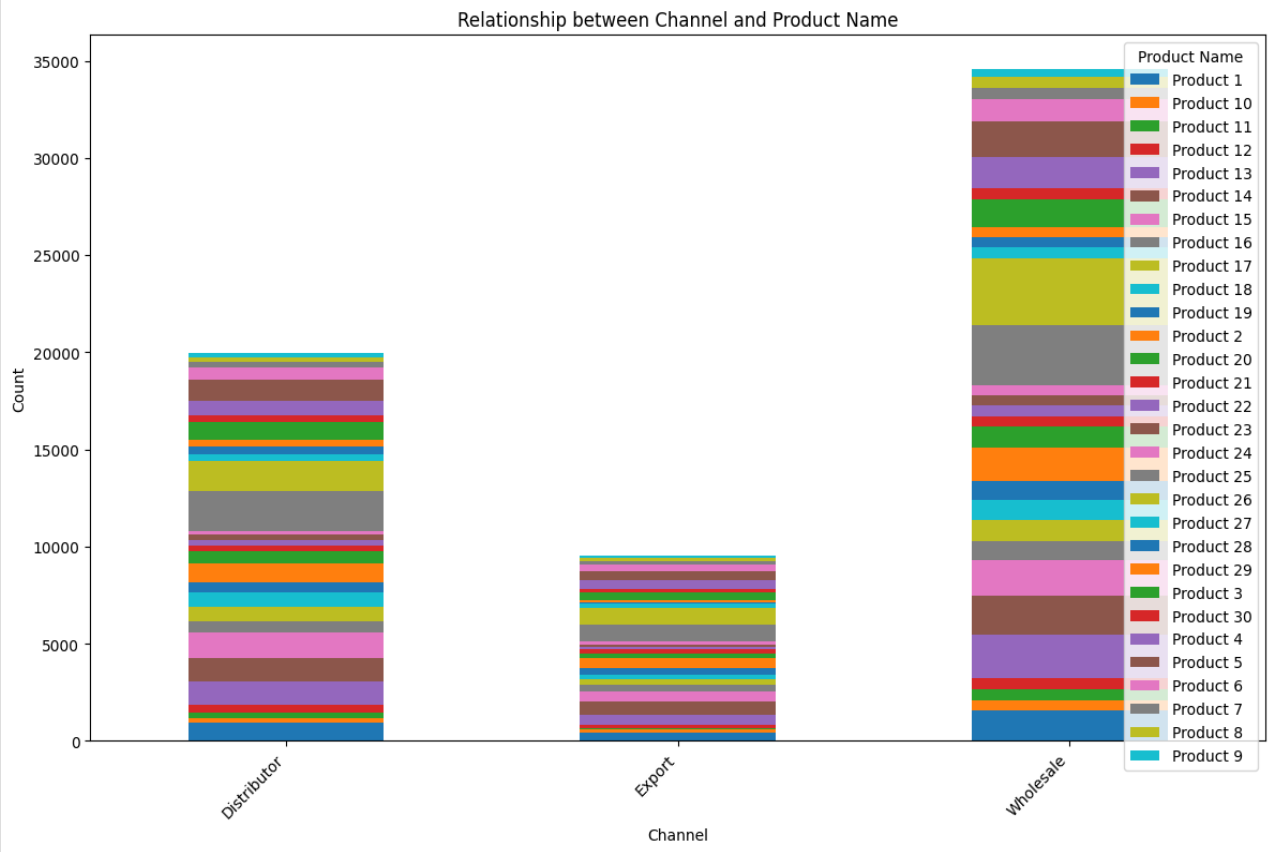
# THE DISTRIBUTION OF ORDERS ACROSS DIFFERENT REGIONS *WITHIN* EACH CHANNEL

This stacked bar plot helps you understand the geographical footprint of each sales channel and how the different channels contribute to the total sales volume in each region. It reveals potential dependencies or strategies where certain channels might be more focused on or successful in particular regions.



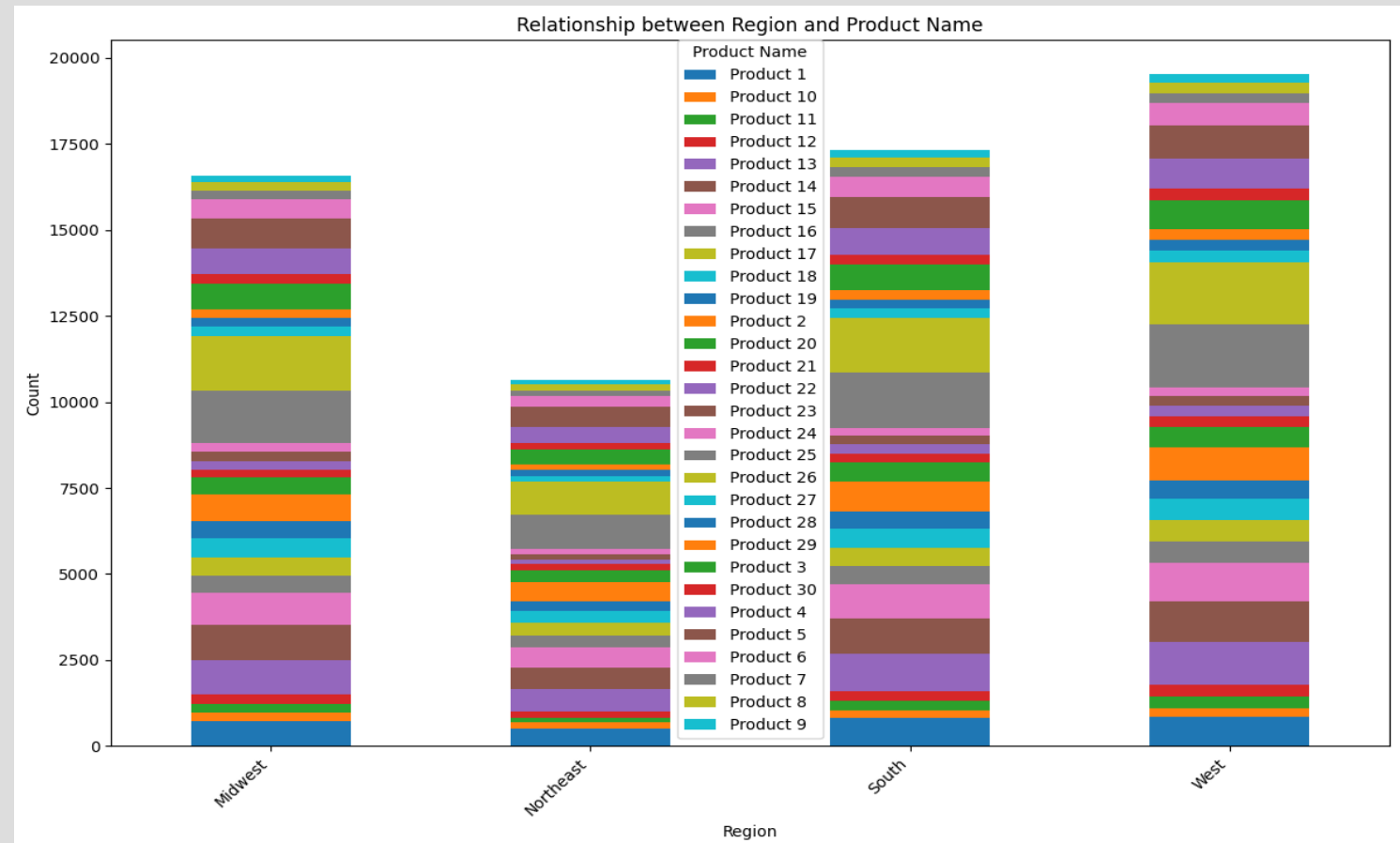
THE DISTRIBUTION OF ORDERS FOR  
EACH PRODUCT ACROSS THE  
DIFFERENT CHANNELS

Provides a visual overview of how product sales are distributed across your different channels. For more specific insights, you would likely need to focus on individual products or groups of products.



## . THE DISTRIBUTION OF ORDERS FOR EACH PRODUCT ACROSS THE DIFFERENT REGIONS

It provides a good visual overview of how product sales are geographically distributed across your regions. You can spot general trends and potentially identify products that perform exceptionally well or poorly in certain regions.



# KEY INSIGHTS

## **Data Overview and Cleaning:**

- The dataset contains sales order information, details about regions, products, customers, state regions, and 2017 budgets.
- Initial data cleaning involved converting date columns and numerical columns with commas to appropriate data types.
- Empty columns were removed from the sales orders data.
- Leading/trailing spaces were removed from categorical columns like customer and product names, and duplicate state region entries were handled.

## **Data Merging:**

- The datasets were successfully merged based on common key columns, creating a comprehensive dataset for analysis.
- New columns for 'Total Cost', 'Profit', and 'Profit Margin' were calculated.
- Missing values in the '2017 Budgets' column were handled by setting values to NaN for orders outside of 2017, which is appropriate as budget data was only provided for 2017.

## **Exploratory Data Analysis (EDA):**

- Numerical Variables:** Descriptive statistics and visualizations (histograms, box plots, density plots) revealed the distribution and potential outliers in key numerical columns such as 'Order Quantity', 'Unit Price', 'Line Total', 'Total Unit Cost', 'population', 'households', 'median\_income', and 'land\_area'.
- Categorical Variables:** Frequency counts and bar plots were used to understand the distribution of orders across different categories like 'Channel', 'Region', 'Product Name', and 'Customer Names'.



- Bivariate Analysis:**

- Numerical-Numerical Relationships:** The correlation matrix and heatmap showed strong positive correlations between 'Unit Price', 'Line Total', and 'Total Unit Cost'. 'Order Quantity' had a moderate positive correlation with 'Line Total'. 'Population' and 'median\_income' showed a weak negative correlation with each other and very weak correlations with the sales-related numerical columns.

- Categorical-Numerical Relationships:** Descriptive statistics and bar plots grouped by categorical variables (Channel, Region, Product Name, Customer Names) revealed how numerical variables like sales and profit metrics vary across different categories. For example, we saw the average profit margin by channel, indicating slight differences in profitability across sales channels.

- Categorical-Categorical Relationships:** Cross-tabulations and stacked bar plots illustrated the relationships and distributions between pairs of categorical columns, such as the distribution of channels within different regions or the frequency of product sales across different customer names.

- Customer and State Performance:** Analysis of top/bottom customers and states by 'Line Total' highlighted the key contributors to sales revenue and identified areas with lower sales.

- Profitability Analysis:** Scatter plots of 'Line Total vs. Profit Margin' and 'Unit Price vs. Profit Margin' showed the relationship between these variables and the distribution of profit margins. The analysis of 'Average Profit Margin by Channel' provided insights into the profitability of different sales channels..

These findings provide a good foundation for understanding the sales data, customer behavior, and profitability across different dimensions

# RECOMMENDATION

- 1.Focus on High-Performing Customers and States:** Identify the characteristics of the top 10 customers and states by Line Total. Develop targeted marketing and sales strategies to nurture these relationships and potentially replicate successful approaches with other customers and in other states.
- 2.Investigate Low-Performing Customers and States:** Analyze the reasons behind the low Line Total in the bottom 10 customers and states. This could involve examining factors like customer churn, low order frequency, or regional economic conditions. Develop strategies to improve performance in these areas, such as customer retention programs or targeted promotional offers.
- 3.Optimize Profitability by Channel:** While the differences in average profit margin by channel are small, further investigation into the cost structure and pricing strategies within the Wholesale channel could identify opportunities to improve its profitability. Consider if adjustments to pricing, product mix, or operational efficiency are needed.
- 4.Deep Dive into Product Performance:** Analyze the performance of individual products, particularly those with high sales volume but lower profit margins, or those with high profit margins but lower sales volume. This could involve examining factors like production costs, pricing, market demand, and competition for specific products
- 5.Explore the Impact of Demographic and Geographic Factors:** The merged data includes demographic information (population, households, median income) and geographic details (latitude, longitude, area code, time zone). Further analysis could explore how these factors influence sales and profitability in different regions and states. For example, are there specific demographic segments or geographic areas that are more profitable?
- 6.Budget vs. Actual Analysis (for 2017):** Since 2017 budget data is available, perform a detailed analysis comparing actual sales and profit against the budget for that year. Identify areas where performance exceeded or fell short of the budget and understand the contributing factors.