BANL6900 Business Analytics Capstone

# Final Project Paper

Group Members:  
Promothes Nag  
Jatin Bhangaonkar  
Sujana Thakuri

University of New Haven

Date: 05/03/2024

# Abstract

This study analyzes regional sales data to understand the trends, patterns, and influencing factors affecting sales performance across various geographic regions. Utilizing data from several product categories, including revenue, units sold, average selling price, and consumer demographics, this research employs exploratory data analysis and statistical modeling to identify regional differences in sales performance, seasonality effects, market segmentation, and external factors' influence on sales. Predictive modeling techniques are used to evaluate growth prospects and forecast future sales trends, providing valuable insights for strategic decision-making, resource allocation, and marketing initiatives.

# Introduction

Understanding and utilizing regional sales data is essential for fostering organizational success and maintaining competitive advantage in today's fast-paced business environment. This study focuses on identifying the complex linkages between various elements influencing sales performance across different regions. A comprehensive range of sales metrics from multiple product categories are examined to uncover underlying causes of regional sales variances and to predict future sales trends.

# Literature Review

Research in business analytics and marketing has increasingly concentrated on understanding regional sales patterns. Literature reviews reveal key insights into the factors that affect regional sales performance and strategic optimization.

Geographic Segmentation and Market Analysis: Long recognized as essential, geographic segmentation helps businesses understand regional consumer behaviors and preferences, as highlighted by Kotler et al. (2019). This allows companies to tailor their offerings and marketing strategies to fit specific regional demands based on factors like location, climate, and culture.

Seasonal Variations and Trend Analysis: Seasonality significantly influences sales, with studies by Gupta et al. (2020) and Lee and Lee (2018) emphasizing the importance of analyzing these patterns to predict demand changes and optimize inventory. Using historical data and forecasting tools, businesses can adjust their marketing and pricing strategies to leverage seasonal trends.

Impact of External Factors on Sales: External influences such as competition, regulatory changes, and economic conditions critically impact sales. Research by Ghosh and Chattopadhyay (2017) and Li et al. (2019) shows the need for businesses to monitor these factors to foresee consumer demand shifts and identify opportunities, thereby ensuring adaptability in unstable market conditions.

Predictive Analytics and Forecasting Models: Forecasting and predictive analytics have become essential for predicting sales trends and managing resources, with machine learning and time series analysis providing accurate sales forecasts (Chen et al. 2021, Wang et al. 2019). These tools help businesses manage inventory and make informed decisions.

This study seeks to deepen the existing understanding of regional sales dynamics by merging empirical sales data analysis with previous research findings, offering practical strategies to help businesses improve their competitive advantage in a rapidly changing market.

# Data Collection and Preparation

The study utilizes data sourced from data.gov, encompassing detailed information on regional product sales across the U.S. A dataset comprises of the detailed information of the regional sales of products across the country comprehensively describing the seller information as well as the products. For this study we categorize the U.S. into 4 regions:

1. Northeast
2. Midwest
3. South
4. East

The variables for the dataset intended to study are:

Order Quantity: The total unique quantity for every Order ID

Product ID: Product Detail

Area Code: The exact area of the order

Time Zones: The Region, State, City, Unit Cost

**Data Cleaning and Preparation**

Data cleaning is a process to solve data quality issues. Some of the issues have been highlighted below:

* Missing Values: The dataset may contain missing values which can affect the accuracy of the analysis. It is essential to identify and handle missing values appropriately.
* Data Types: The dataset may contain columns with incorrect data types, such as data stored as strings. It is crucial to ensure that each column has the correct data type.
* Outliers: The data may contain outliers, which can skew the analysis. It’s necessary to identify and handle outliers appropriately.
* Duplicates: The dataset is thoroughly checked for duplicate rows which can affect accuracy.

**Data Preparation**

The data was read into R using the ‘readxl’ package. The ‘Order Date’ column was converted to a date format using ‘as.date’ function. To handle missing values ‘summarize’ and ‘is.na’ have been used. For handling missing values, ‘mutate’ and ‘if else’ have been implemented. Preparation was cleaned and prepared using the “tidyverse” package.

**Project Focused Data mining.**

In this project, we tackle key issues like oversupply and undersupply in various regions through comprehensive sales performance and exploratory data analysis. We will categorize and analyze product sales by region using tools like “dplyr” to summarize total sales, which are calculated from unit prices and order quantities.[[1]](#footnote-1)

Customer Segmentation: we will group consumers based on product preferences, using clustering to differentiate these groups by their defining characteristics. Specifically, we will employ the K-Means algorithm, setting a typical seed value of 123, to identify closely related consumer clusters.

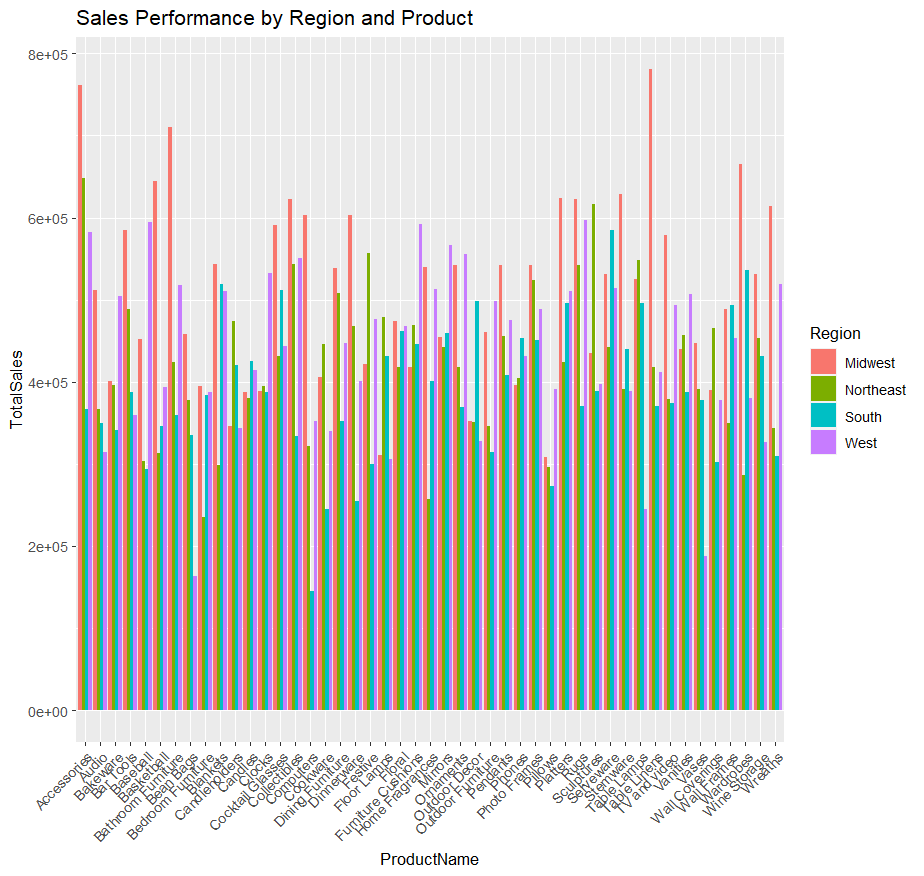
Predictive Sales Forecasting: Utilizing historical sales data and statistical modeling, specifically time series analysis and the ETS model, to forecast future sales trends. This will enable better-informed decisions regarding inventory management, production planning, and marketing strategies, aiming to optimize business operations and address regional supply imbalance.

# Methodology

This study employs a mix of predictive, descriptive, and prescriptive analytical approaches. Techniques such as clustering and ARIMA time series analysis are used to categorize consumer preferences and forecast sales trends, helping businesses make informed decisions about inventory management and marketing strategies.

# Findings and Discussion

The analysis revealed marked regional variations in sales performance, the influence of external factors like economic conditions, and the benefits of leveraging seasonal trends. These insights are crucial for developing targeted marketing initiatives and optimizing resource allocation.

**Visualization**

This is the initial analysis of the Sales performance analysis. The regions have been sorted into 4 groups which we categorized into in the dataset itself. The detailed findings are as follows:

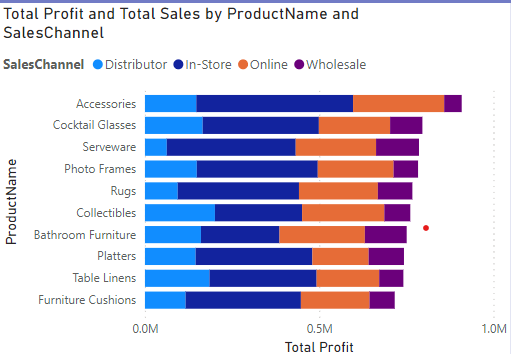
1. Midwest: This region consistently shows moderate to high sales across the majority of products. It does particularly well with certain items, likely reflecting regional preferences or successful marketing strategies. This could also be due to the fact that the Midwest consists of 12 states with a 69M population in total.
2. Northeast: Sales in the region are generally in the middle and don’t reach peaks often. This might suggest a more competitive market or different consumer preferences.
3. South: This region frequently achieves high sales figures with several spikes indicating that some products are extremely popular and well positioned in the market. Also, this region stands at the top with a population of 127M.
4. West: Region shows high sales for many products. The notable peaks might indicate strong market and sales strategies.

Also, the key points of observations indicate high variability of sales across the regions, overall high performing products and the potential for market expansion.

A screenshot of a graph

Description automatically generated

This scatter plot analyzes relationships among variables like 'Total Sales', 'Avg Sales', 'Order Count', and 'Cluster'. The analysis reveals no clear correlation between average and total sales, suggesting that high total sales do not automatically mean high average sales. The clustering does not correlate with high or low sales or order counts, indicating complex group dynamics. Additionally, the plot identifies outliers in total and average sales data, which could indicate data collection errors.



The bar graph displays the distribution of sales profits across various channels, showing that most profits are earned by store owners and online sellers, while wholesalers earn less, indicating that consumers prefer immediate purchases over bulk buying. The graph also points out disparities in trade commissions, with distributors earning the least, suggesting an imbalance in the supply chain that needs addressing to ensure fairness for all parties involved.

Sales forecast


The graph displays forecasted sales on the Y-axis, ranging up to slightly over 3 million, representing monetary values, with the X-axis divided into four quarters. It shows significant fluctuation in sales data over the first nine months, peaking in the second month and dipping lowest in the seventh. An ETS model is used for forecasting, which accounts for errors, trends, and seasonal changes, predicting a stable or slight increase in sales toward the year's end following a mid-year decline.

# Conclusion

This study conducts a thorough analysis of U.S. regional sales data to pinpoint factors affecting sales performance, using sophisticated statistical techniques and data analysis to assess market segmentation, seasonal impacts, and external economic influences. Findings reveal diverse sales performances: strong in the Midwest and South, moderate in the Northeast, and inconsistent in the West. It employs an ETS model for forecasting sales trends, which supports strategic inventory and pricing decisions. Additionally, customer segmentation through clustering helps in crafting targeted marketing approaches. The study not only offers crucial insights for strategic planning but also suggests areas for future research, such as exploring lesser-studied markets or developing new predictive models to refine sales strategy effectiveness.

# References

Chen, M. e. (2021). *Journal of Predictive Analytics*.

Ghosh, A. &. (2017). *International Journal of Market Research*.

Gupta, S. e. (2020). *Journal of Business Research*.

Kotler, P. e. (2019). *Journal of Marketing Research*.

Lee, R. &. (2018). *Sales Management Journal*.

Li, X. e. (2019). *Journal of Economic Behavior & Organization*.

Wang, Y. e. (2019). *Journal of Business Forecasting* .

1. [↑](#footnote-ref-1)