Insights Unveiled: Data Analysis and Sentiment Analysis for Business Advancement

Zarina Mam, Bhushan Sonsale, Prafull Sonawane, Avadhoot Sutar, Tanmay Agarkar, Sumit Tambe
Department of Artificial Intelligence & Data Science
Vishwakarma Institute of Technology
Pune, India
{zarina.mam, bhushan.sonsale21, prafull.sonawane21, avadhoot.sutar21, tanmay.agarkar21, sumit.tambe}@vit.edu

***Abstract:* This research paper delves into comprehensive sales and sentiment analysis using Mama Earth Company's sales data for the years 2022-23. The study encompasses various parameters such as product categories, states, pack sizes, and products, along with sentiment analysis of customer reviews. Utilizing advanced tools like Plotly for visualization and Streamlit for the interface, the analysis provides valuable insights into consumer behavior, market trends, and opportunities for business optimization. The paper also discusses the technical stack and methodology employed in the analysis, offering a detailed overview of the research process. Through in-depth statistical analysis and visualization techniques, the paper presents actionable insights and solution recommendations for enhancing business performance and customer satisfaction.**

Keywords — Sales analysis, Sentiment analysis, Mama Earth Company, Data visualization, Consumer behavior, Market trends, Business optimization.

I. Introduction

Every institution's annual sales data is a large dataset with many variables to take into account while doing sales analysis. In addition, we must exclude any blank spots from the dataset. Customer reviews make up a column in some datasets, and we also need to analyse their sentiments. Because working at a high level can lead to missing crucial details that could result in business losses, it's critical to have a visualized data analysis that takes into account all relevant parameters and sub parameters in addition to customer review sentiment analysis. For this study, we have chosen the dataset from Mama Earth Company's sales data for the years 2022-23. This dataset includes essential parameters such as product name, price, product category, review date, pack size, review count, states, review content, and several others. Every Dataset analysis needs some technical tools lets understand them.

Four parameters—Product Category, States, Pack Size, and Products—as well as an Overall Analysis component have been taken into consideration for the data analysis. Users have the ability to choose any criteria, including price, product category, and others, under the section titled "Overall Analysis." When price is chosen, for instance, a percentage table displaying the price-wise analysis in ascending order is produced. This table shows information such as which goods, priced at 399, had the most sales (20.2395%). This method offers a thorough understanding of how each aspect affects sales. Moreover, comprehension of information is improved by the use of visuals like pie charts and line graphs. Sales trends and patterns are clearly displayed by these graphic tools. Additionally, we perform Sentiment Review Comparison using Natural Language Processing (NLP) to categorize reviews into positive and negative sentiments. Graphs based on this sentiment analysis are plotted to identify areas of improvement, particularly focusing on products with weaker performance.

We've incorporated some essential graphs and plots for analysis. However, for those unfamiliar with statistics, we've introduced a helpful feature. Upon clicking the 'Get Insight' button, users receive a summary of business analysis and suggested solutions. Here's how it functions: We present a data table, and your task is to analyse it from a business standpoint. We focus on two key labels: 'Feature1' (the aspect under analysis) and 'Number of Sold Products at That Price' (our sales data). Your insights should follow this format:

1. Insight Point 1: <Insight description>

2. Insight Point 2: <Insight description>

For each insight, provide at least four analysis points and 2-3 solution suggestions. We utilize LLM (Large Language Model) and the Google Gemini API to fetch insights directly from the data.

When you get statistical Analysis of any Data, it becomes easier to find out insights out of that data, future planning becomes easier and also you become cautious for furthermore upcoming risks, you stop investing your money on the products which are not sold as you want. The Interface of analysis is made in Streamlit Library of Python. For graphs and visualization purposes Plotly library is used. Also, for Sentiment Analysis, Natural language Processing & process of Tokenization is in use.

II. Literature Review

The paper reviews various word and sentence semantic similarity techniques, proposing a model to compute accuracy in Twitter datasets. Word techniques include corpus-based, knowledge-based, and feature-based; sentence techniques include string and set-based, word order-based, POS-based, and syntactic dependency-based. Atish's measure outperforms others in real-world Twitter dataset evaluations. [1]. The paper addresses sentiment analysis for customer review classification in the context of the expansive World Wide Web. It preprocesses datasets, extracts meaningful adjectives as feature vectors, and applies machine learning algorithms (Naive Bayes, Maximum Entropy, SVM) alongside Semantic Orientation based WordNet. Performance evaluation is conducted in terms of recall, precision, and accuracy to enhance decision-making processes [2].

The paper investigates sentiment analysis on large Amazon Fine Food review datasets, employing Apache Spark's data processing system. Methods like Linear SVC, Logistic Regression, and Naïve Bayes, implemented through MLlib, achieve over 80% accuracy. Linear SVC demonstrates superior efficiency compared to Naïve Bayes and logistic regression [3].

The research paper focuses on sentiment analysis across various topics using supervised machine learning, particularly examining opinions from humanity to terrorism. Preprocessing involves converting unstructured reviews into structured format, numerical representation, and sentiment scoring. Support Vector Machine and Naïve Bayes algorithms are compared, with SVM demonstrating superior accuracy in classifying airline reviews as positive or negative[4].

The paper investigates sentiment analysis on big data using the Naïve Bayes algorithm and MapReduce framework. It addresses the challenge of efficiently analyzing vast social network data. Preprocessing techniques and linguistic methods are applied to Twitter data, resulting in a 5% enhancement in sentiment analysis accuracy, achieving 73% accuracy on the Stanford Sentiment dataset [5]. The paper introduces a novel semantic analysis-driven model for extracting relevant data from vast online repositories. It comprises semantic similarity-based feature selection, data summarization, and a deep neural network classifier. The model addresses existing limitations by enhancing relevancy accuracy through innovative techniques. Experimental validation demonstrates its effectiveness in retrieving relevant data from internet resources, bolstered by testing with recognized datasets [6].

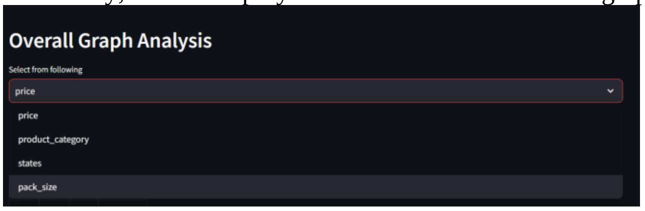
This paper reviews eight publicly available datasets for Twitter sentiment analysis evaluation, highlighting a common limitation: the lack of distinct sentiment annotations for tweets and their entities. To address this, the authors introduce STS-Gold, a new dataset with individual annotations for tweets and targets. Additionally, they conduct a comparative analysis of dataset characteristics and their impact on sentiment classification performance [7]. This paper introduces a novel method for sentiment analysis on Twitter by incorporating semantic features alongside traditional features. Entities extracted from tweets are associated with their semantic concepts, enhancing sentiment prediction. Results demonstrate an average F harmonic accuracy increase of around 6.5% and 4.8% compared to baseline methods. Comparison with sentiment-bearing topic analysis reveals semantic features' superiority in various sentiment classification metrics [8]. This paper discusses the importance of sentiment analysis in decision-making processes, highlighting issues such as polarity shift and data sparsity. Despite the introduction of various methods, existing machine learning algorithms like Naïve Bayes and Support Vector Machine have limited effectiveness in sentiment classification. The survey reviews different sentiment analysis methodologies and approaches, providing insights into their strengths and limitations, culminating in a comparative analysis of addressed issues and metrics used [9]. This paper explores the integration of semantic analysis into Industry 4.0 practices, aiming to automate fault detection and resolution in machines. Leveraging case-based reasoning, the system analyzes reports from operators to propose solutions when faults occur. This approach significantly reduces the need for manual intervention, streamlining the repair process and improving efficiency by providing operators with relevant historical repair data [10].

This survey provides valuable insights into the current state of sentiment analysis research, paving the way for future investigations to bridge existing gaps and propel the field toward even greater advancements.

III. Methodology

A. Project Working

In our research paper, we've organized the analysis into five distinct sections: Overall, Product Category, States, Pack size, and Products. When delving into the Overall Graph Analysis segment, users are prompted to select a singular parameter for conducting the overarching analysis. For instance, they might opt to scrutinize sales data based on price, product category, or any other pertinent criterion. To illustrate these analyses effectively, we employ visual aids such as graphs.



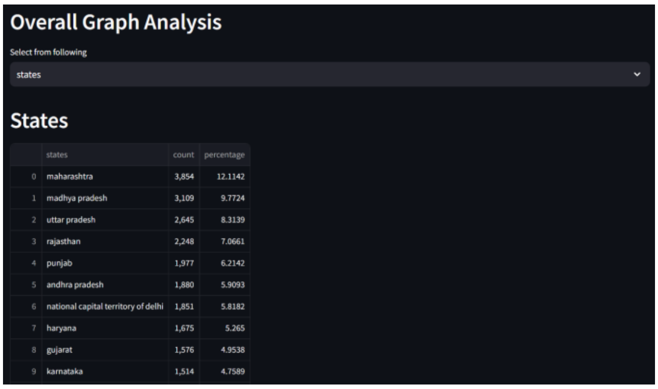
Overall Graph Analysis

Select from following

price

Fig 1.1

When a specific parameter is selected, users receive comprehensive analysis related to that parameter. For instance, if "States" is selected, the system provides an in-depth analysis focusing on geographical states. Sales Count of Every State



Overall Graph Analysis

Select from following

states

States

	states	count	percentage
0	maharashtra	3,854	12.1142
1	madhya pradesh	3,109	9.7724
2	uttar pradesh	2,645	8.3139
3	rajasthan	2,248	7.0661
4	punjab	1,977	6.2142
5	andhra pradesh	1,880	5.9093
6	national capital territory of delhi	1,851	5.8182
7	haryana	1,675	5.265
8	gujarat	1,576	4.9538
9	karnataka	1,514	4.7589

Fig 1.2

Figure 1.2 visually demonstrates the process of selecting the "state" parameter and subsequently obtaining a sales analysis for that state, presented in ascending order. For instance, the analysis may reveal that Maharashtra State boasts the highest sales percentage, approximately 12.1142%. This insight highlights the value of analysis in simplifying complex data exploration tasks, which may be challenging to accomplish directly by searching through a CSV file. Furthermore, Figure 1.3 complements the analysis with additional visual aids, including a pie chart and a line graph, enhancing the understanding and interpretation of the data presented in the analysis.

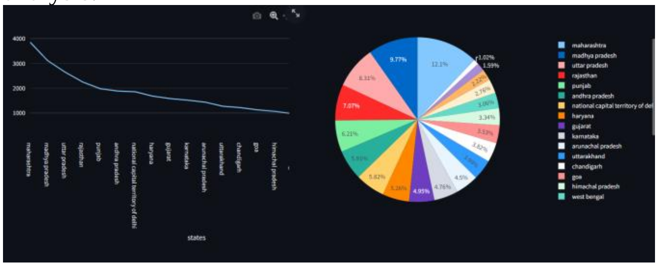


Fig 1.3

From Figure 1.3, it becomes apparent which states require improvement in sales, as seen with Himachal Pradesh having the lowest sales among all states. This insight aids in directing efforts towards areas needing sales enhancement.

In addition to sales analysis, our research paper includes a section on state-wise sentiment analysis. This analysis employs a binary distribution of reviews, categorizing them into positive and negative sentiments. Figure 1.4 visually represents this sentiment analysis, providing a clear understanding of customer sentiments.

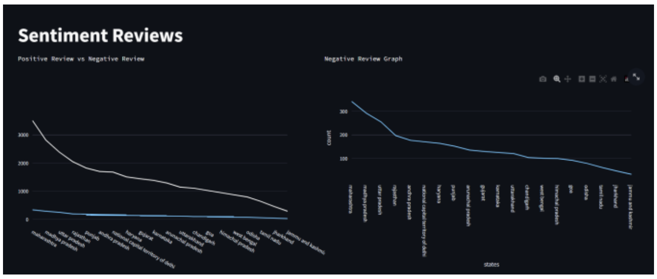


Fig 1.4

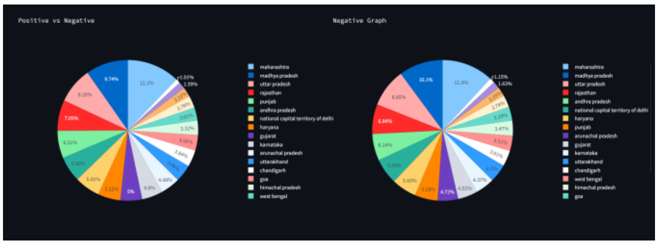
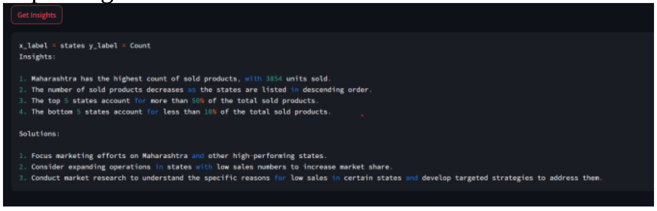


Fig 1.5

Analyzing the sentiments from Figures 1.4 and 1.5, we observe that Maharashtra has the highest percentage of positive reviews, indicating a strong positive sentiment among consumers in that state, around 11.9%. This insight suggests that products are well-received and loved by consumers in Maharashtra.

However, it's notable that Maharashtra also has the highest percentage of negative reviews. This finding underscores the importance of further improvement efforts in Maharashtra to address any issues highlighted by negative feedback. By enhancing product offerings or addressing customer concerns, there's potential to boost sales further in this region. This analysis emphasizes the significance of understanding customer sentiments at a granular level, enabling targeted strategies for improving sales and overall customer satisfaction.



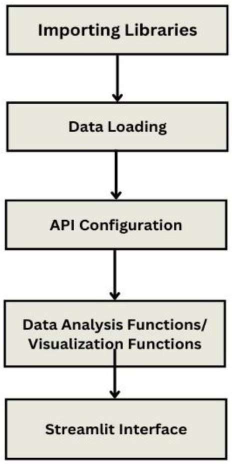
```
x_label = states y_label = Count
Insights:

1. Maharashtra has the highest count of sold products, with 3854 units sold.
2. The number of sold products decreases as the states are listed in descending order.
3. The top 5 states account for more than 50% of the total sold products.
4. The bottom 5 states account for less than 10% of the total sold products.

Solutions:

1. Focus marketing efforts on Maharashtra and other high-performing states.
2. Consider expanding operations in states with low sales numbers to increase market share.
3. Conduct market research to understand the specific reasons for low sales in certain states and develop targeted strategies to address them.
```

Fig 1.6

We've incorporated a feature were tapping on "Get Insights" provides access to all insights derived from the analysis. We utilized LLM (Large Language Model) and the Google Gemini API to extract information from graphs, ensuring a comprehensive understanding of the data. This analysis is available for each parameter under the Overall Analysis section, including price, product category, states, and pack size. Users can access detailed insights and a brief description of each parameter's analysis by utilizing the "Get Insights" feature, facilitating a deeper comprehension of the data and its implications.

B. *Technical Working*

The technical stack behind this project includes the Plotly library for visualizing different types of graphs such as line plots, bar charts, and pie charts. Additionally, the Streamlit library is utilized for building the user interface, enabling interactive features like dropdowns, buttons, and data tables. For a comprehensive overview, refer to diagram 1.6.



Importing Libraries

Data Loading

API Configuration

Data Analysis Functions/ Visualization Functions

Streamlit Interface

Fig 1.6

- **Library Setup**

The research project begins by importing essential libraries required for data handling, analysis, visualization, and interface development. These libraries include requests, BeautifulSoup, pandas, numpy, plotly, matplotlib, streamlit, and google.generativeai.

- **Data Loading**

Data files, namely clean_reviews.csv and sentiment_reviews.csv, are loaded and preprocessed using the panda's library. This step ensures that the data is in a suitable format for analysis and visualization.

- **API Configuration**

The Google Gemini API key, stored in an environment variable (os.environ['GEMINI_API_TOKEN']), is configured to access external data sources and enhance the functionality of the research project.

- **Function Definitions**

Custom functions are defined to perform various tasks within the project:

I. get_insight: Generates insights based on data analysis and prompts using the Google Gemini API.

II. product_img_url: Fetches image URLs of products for display in the interface.

III. get_line, get_bar, get_pie: Functions to create line plots, bar charts, and pie charts using the Plotly library for data visualization.

- **Streamlit Interface**

The Streamlit library is utilized to create an interactive user interface (UI) for the research project. The interface includes features such as sidebar radio selectors, dropdowns, buttons, and data tables for user interaction.

3. Results and Discussions

Price vs. Sales Volume:

The analysis reveals a clear inverse relationship between price and sales volume, indicating that consumers are more inclined to purchase products at lower prices. This aligns with established economic principles, where demand typically decreases as price increases. Businesses should consider this relationship when setting prices to maximize revenue and maintain competitiveness in the market.

- Price Ranges:

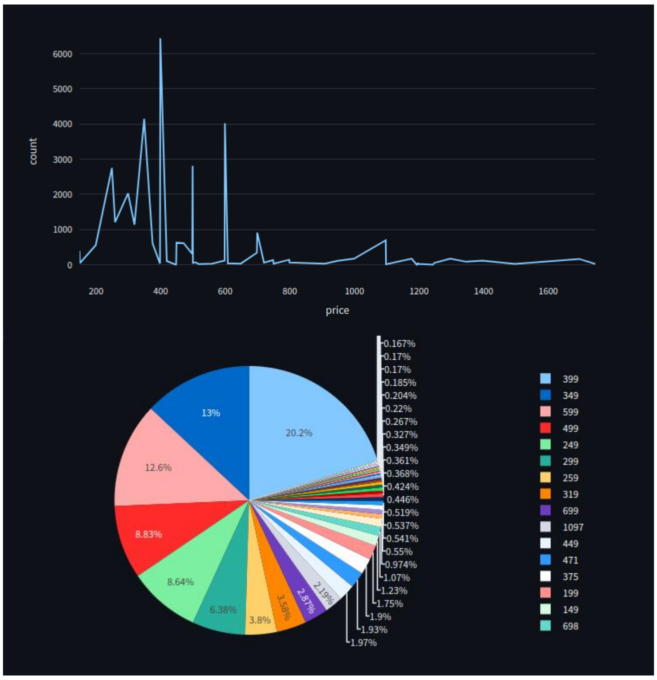
Certain price ranges, particularly between $1,000 and $2,000, stand out for their higher sales volumes. This suggests that products within these price brackets are more attractive to consumers. Businesses can capitalize on this by focusing marketing efforts on products within these ranges or by adjusting pricing strategies to align with these popular price points.

- Outlier Prices:

The presence of outlier prices with significantly higher sales volumes indicates that promotional activities or discounts may have a substantial impact on consumer behavior. Businesses should monitor these outliers closely and consider replicating successful strategies to drive sales and attract customers.

- Seasonal Trends:

The analysis hints at potential seasonal trends in sales volume, with certain prices being more popular during specific times of the year. Understanding these seasonal variations can help businesses adjust their marketing and pricing strategies to capitalize on seasonal demand fluctuations.

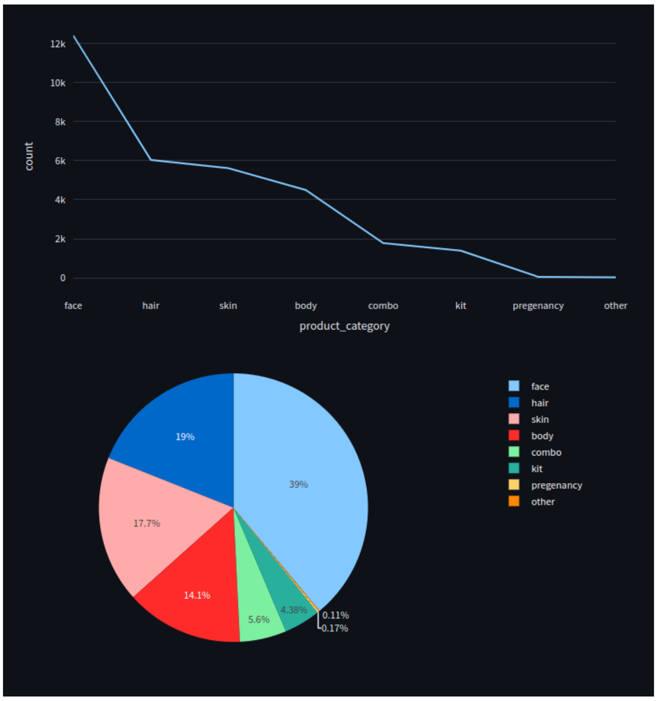


Business Insights from Product Categories:

The analysis of product categories reveals valuable insights into consumer preferences and market trends. The high demand for face products suggests a strong market for skincare and cosmetic products aimed at enhancing facial appearance. Additionally, the significant market share held by hair and skin products highlights the importance of personal grooming and self-care in consumer behavior.

- Solution Recommendations for Product Categories:

Based on the insights from the product categories analysis, businesses can develop targeted strategies to capitalize on consumer preferences and market trends. Investing in marketing efforts to target face product customers can help maximize revenue from this high-demand category. Expanding hair and skin product offerings can also help businesses cater to diverse customer needs and preferences, potentially increasing market share and profitability. Developing high-value combination products can further appeal to consumers looking for convenience and effectiveness in their skincare routines. Lastly, exploring the untapped niche market for pregnancy-related products presents an opportunity for businesses to differentiate themselves and capture a specialized segment of the market.

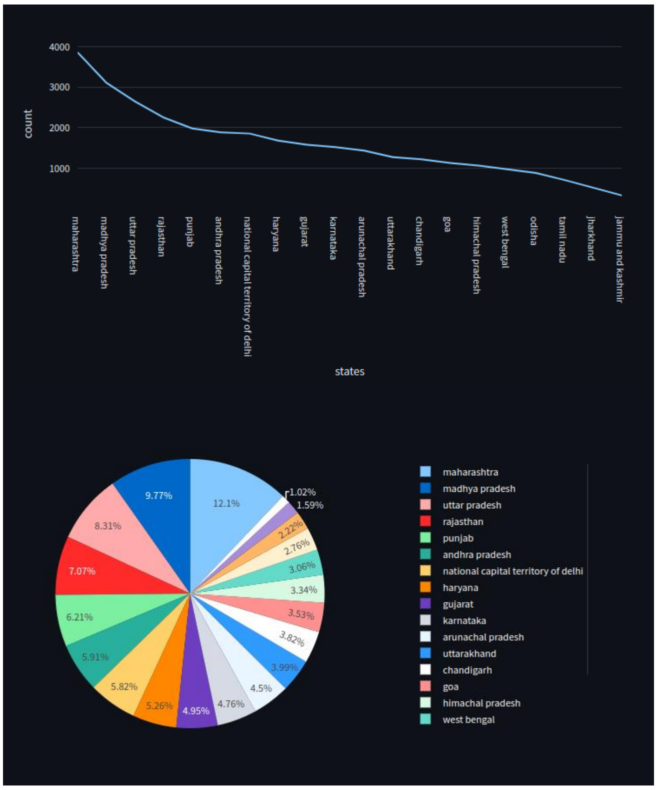


Insights from Regional Sales Analysis:

The analysis of regional sales highlights key insights into market presence and potential for growth across different states. Maharashtra emerges as the state with the highest number of sold products, indicating a strong market presence. Other states like Madhya Pradesh, Uttar Pradesh, Rajasthan, and Punjab also show significant sales, suggesting competition in these regions. On the other hand, states like Arunachal Pradesh, Uttarakhand, Chandigarh, and Goa exhibit lower product sales, indicating a need for targeted strategies to improve market accessibility and demand generation in these areas.

- Solutions for Regional Sales:

To capitalize on the high demand and competitive landscape in states like Maharashtra, Madhya Pradesh, Uttar Pradesh, Rajasthan, and Punjab, businesses should prioritize marketing and sales efforts in these regions. Exploring strategies to enhance brand visibility and distribution channels in states like the National Capital Territory of Delhi and Haryana can help capture a larger market share. Additionally, conducting market research and tailoring marketing campaigns to specific regions like Arunachal Pradesh, Uttarakhand, Chandigarh, and Goa can increase product awareness and drive sales in these under-performing areas.

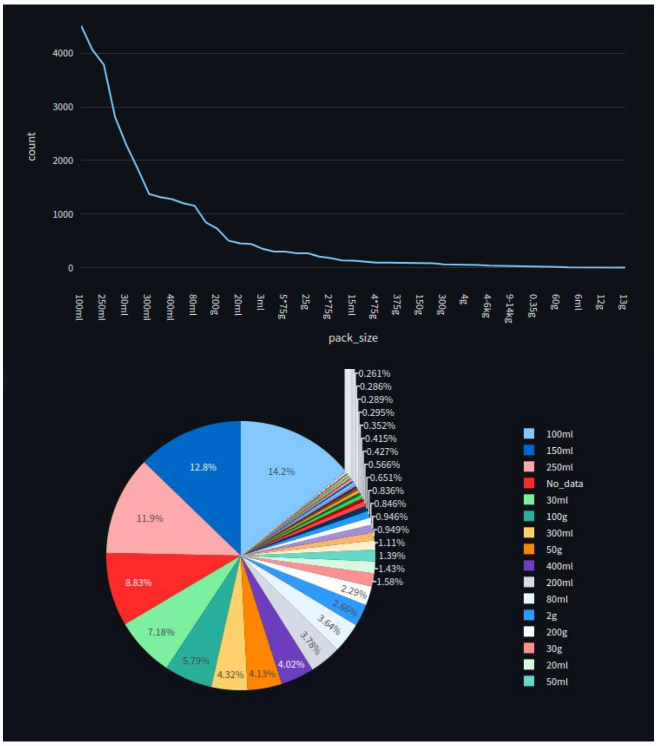


Insights from Pack Size Analysis:

The analysis of pack sizes reveals consumer preferences and market trends related to packaging. The top-selling pack sizes, such as '100ml', '150ml', and '250ml', should be prioritized in production and marketing campaigns to meet consumer demand. The trend towards smaller pack sizes suggests that consumers value convenience, portability, and reduced waste. However, under-represented pack sizes like '2g', '3ml', '6ml', '8ml', '12g', and '13g' present opportunities for businesses to fill gaps in the market and cater to evolving consumer preferences.

- Solutions for Pack Size Optimization:

To optimize production and inventory for popular pack sizes, businesses should ensure an adequate supply of the top-selling pack sizes to meet demand and minimize stockouts. Developing innovative packaging solutions for smaller sizes can make them more appealing to consumers seeking convenience and portability. Introducing new pack sizes based on gaps in the market can help businesses address under-represented categories and align with consumer preferences for specific pack sizes.



4. Conclusion and Future Scope

In conclusion, the analysis of price, product categories, regional sales, and pack sizes provides valuable insights into consumer behavior and market trends. By leveraging these insights, businesses can develop targeted strategies to optimize pricing, product offerings, regional sales, and packaging, ultimately maximizing revenue and profitability.

References

[1] Belal Abdullah Hezam Murshed, Hasib Daowd Esmail Al-ariki, Suresha Mallappa. " Semantic Analysis Techniques using Twitter Datasets on Big Data: Comparative Analysis Study" IEEE transactions on pattern analysis and machine intelligence 35, no. 4 (2012): 882-897.

[2] Geetika Gautam, Divakar Yadav. " Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis " IEEE Access 9 (2020): 1420-1427.

[3] Hafiz Muhammad Ahmed, Mazhar Javed Awan, and Awais Yasin." Sentiment Analysis of Online Food Reviews using Big Data Analytics" In 2017 13th International Conference on Semantics, Knowledge and Grids (SKG), pp. 33-37. IEEE, 2017.

[4] Abdul Mohaimin Rahat, Abdul Kahir, Abu Kaisar Mohammad Masum " Comparison of Naive Bayes and SVM Algorithm based on Sentiment Analysis Using Review Dataset " In Proceedings of the Intelligent Vehicles' 95. Symposium, pp. 258-263. IEEE, 1995.

[5] Mariam Khader, Arafat Awajan, and Ghazi Al-Naymat. " The Impact of Natural Language Preprocessing on Big Data Sentiment Analysis" IEEE Transactions on Intelligent Transportation Systems 22, no. 6 (2021): 3234-3246.

[6] Antony Rosewelt, Arokia Renjit. " Semantic analysis-based relevant data retrieval model using feature selection, summarization and CNN" In 2019 international conference on communication and signal processing (ICCSP), pp. 0157-0160. IEEE, 2019.

[7] Hassan Saif, Miriam Fernandez, Yulan He and Harith Alani. " Evaluation Datasets for Twitter Sentiment Analysis." In 2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA), pp. 262-267. IEEE, 2021.

[8] Hassan Saif, Yulan He and Harith Alani. " Semantic Sentiment Analysis of Twitter." In 2021 5th international conference on intelligent computing

and control systems (ICICCS), pp. 954-959. IEEE, 2021.

[9] Ms.A.M.Abirami, Ms.V.Gayathri. " A SURVEY ON SENTIMENT ANALYSIS METHODS AND APPROACH." In 2018 International Conference on Intelligent and Innovative Computing Applications (ICONIC), pp. 1-4. IEEE, 2018.

[10] Alberto Rivas, Luc´ıa Mart´ın, In´es Sitt´on1 Pablo Chamoso, Javier J. Mart´ın-Limorti, Javier Prieto, and Alfonso Gonz´alez-Briones. " Semantic analysis system for Industry 4.0." In 2017 International Conference on Machine Learning and Cybernetics (ICMLC), vol. 1, pp. 156-161. IEEE, 2017.