Experiment 1

Aim: Case Study Framing & Dataset Preparation

Objective: Define a real-world domain problem, benchmark existing solutions, acquire data (API/scraping/open portals). Document versioning plan with DVC.

Problem Statement

Company Overview

The Whole Truth Foods is an Indian clean-label food brand founded in 2019 that focuses on creating transparent, healthy snacks and food products. The company positions itself as an alternative to conventional processed foods by using minimal, natural ingredients and avoiding artificial additives, preservatives, and refined sugars.

Problem Definition

Domain Problem: Consumer Preference Prediction for Clean-Label Food Products in the Indian Market

Company Background: The Whole Truth Foods is an Indian clean-label food brand founded in 2019 that creates transparent, healthy snacks using minimal natural ingredients while avoiding artificial additives, preservatives, and refined sugars. The company has positioned itself as a premium alternative to conventional processed foods in India's rapidly growing health-conscious consumer segment.

Business Challenge: The Indian clean-label food market is experiencing unprecedented growth, with consumers increasingly prioritizing health and transparency post-COVID-19. However, The Whole Truth Foods faces critical challenges in predicting which product attributes will drive consumer adoption and market success. The company struggles with understanding consumer sentiment across product categories, identifying optimal pricing strategies, and forecasting demand patterns for effective inventory management.

Research Objectives: This study aims to develop a data-driven framework to predict consumer preferences for clean-label food products by analyzing social media sentiment, e-commerce behavior, and market positioning. This aims to create predictive models to identify key factors influencing consumer choice, benchmark. The Whole Truth Foods against competitors, and uncover market gaps for new product development opportunities.

Key Research Questions:

1. Which product attributes most strongly influence customer ratings and purchase likelihood for clean-label food items?
2. How can consumer sentiment scores (derived from reviews) for The Whole Truth Foods be predicted and compared with competitors across available channels in the dataset?
3. Can machine learning models accurately predict the commercial success of new products using composition, pricing, and historical data?
4. What seasonal and demographic patterns can be identified in clean-label food consumption to improve sales forecasting and demand planning?

Expected Business Impact: The research will provide actionable insights for product development, pricing optimization, and marketing strategy. Success will be measured through sentiment analysis accuracy (≥85%), consumer preference classification recall (≥95%), and demand prediction RMSE (≤15%). The deliverables will include predictive models, competitive analysis dashboard, and strategic recommendations for market expansion.

Data Sources: Multi-channel data collection from social media platforms (Twitter, Instagram, Facebook), e-commerce sites (Amazon, Flipkart), company websites, and industry reports. The integrated dataset will enable comprehensive analysis of consumer behavior, market trends, and competitive positioning in the Indian clean-label food ecosystem.

Benchmark Analysis

Existing Solutions in Food Industry Analytics

1. Traditional Market Research Approaches:

- Nielsen Consumer Insights
- Euromonitor Market Reports
- Manual surveys and focus groups
- Limitations: High cost, limited real-time insights, small sample sizes

2. Digital Analytics Solutions:

- Social media sentiment analysis tools (Brandwatch, Hootsuite)
- E-commerce analytics platforms (Jungle Scout for Amazon)
- Google Trends analysis
- Limitations: Platform-specific insights, lack of integration across channels

3. Academic/Open-Source Approaches:

- VADER sentiment analysis for social media
- Collaborative filtering for product recommendations
- Time series forecasting using ARIMA/Prophet
- Limitations: Generic models not tuned for food industry specifics

Proposed Improvements

Our approach will integrate multiple data sources and apply ensemble methods to:

- Leverage combined structured (product id,name,ingredients) and unstructured (reviews, social media) data with actual purchase behavior
- Use computer vision for packaging and ingredient analysis
- Use advanced NLP techniques and regression models to link sentiment trends with sales fluctuations.
- Implement real-time data pipeline for continuous monitoring
- Focus on maximizing recall of negative sentiment to proactively address customer issues.

Data Acquisition Strategy

For this project, all data was sourced from **The Whole Truth Foods** website using a hybrid data acquisition approach:

- **Automated Extraction**: Browser-based scraping tools and extensions (such as the **Instant Data Scraper Chrome Extension**) were used to extract structured product-related data like product names, prices, and descriptions.

- **Manual Scraping**: In cases where automated tools were unable to retrieve certain details—such as customer reviews or specific nutritional information—data was collected manually to ensure completeness.

- **Synthetic Data Generation**: To supplement areas where real user-generated content (like reviews or user metadata) was limited or unavailable, the **Faker** library was used to generate realistic synthetic data. This helped simulate a more comprehensive dataset suitable for downstream Natural Language Processing (NLP) tasks such as summarization and classification.

This combined approach ensured that the dataset was both rich in content and representative of real-world product and user interaction data, enabling effective analysis and model development.

Description of Dataset

The dataset consists of **10,000 records** and **31 attributes**, collected from **The Whole Truth Foods** website using browser scraping tools, manual methods, and synthetic data generation (via the Faker library) to enrich fields like reviewer information.

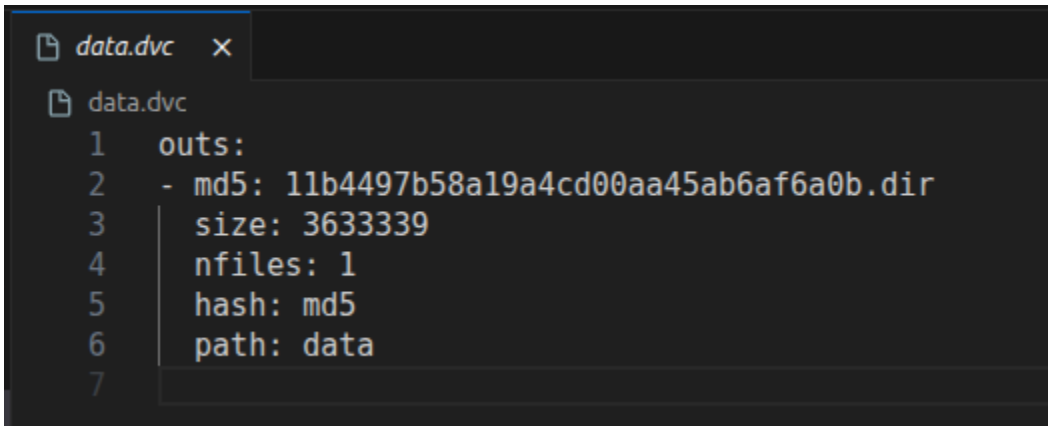
- **product_id**: Unique identifier for each product (e.g., `TWT_001`). Helps track and link reviews, sales, and other product-specific data.
- **product_name**: Official product name from the website (e.g., "Double Cocoa Mini Protein Bars - Box of 12").
- **brand**: The brand associated with the product (in this dataset, always "The Whole Truth").
- **category**: Product category such as Mini Bars, Protein Powder, Muesli, etc.
- **ingredients**: Comma-separated list of ingredients (e.g., "Almonds, Cocoa Powder, Protein Powder").
- **clean_label**: Indicates whether the product is marketed as "clean label" (Yes = no artificial additives).
- **price**: Selling price in Indian Rupees before discounts.
- **discount**: Discount percentage applied (e.g., 15 = 15%).
- **units_sold**: Number of product units sold. Useful for demand analysis.
- **average_rating**: Mean customer rating (1.0 to 5.0).
- **num_reviews**: Total number of reviews for the product.
- **shelf_life**: Shelf life after manufacturing (e.g., "6 months", "12 months").
- **packaging_type**: Type of packaging—jar, pouch, box, etc.
- **review_id**: Unique identifier for each review (e.g., `RVW_0001`).
- **review_text**: Text content of the customer review (primary data for NLP).
- **rating**: Individual review rating (1 to 5 stars).
- **platform**: Platform where the review was posted (e.g., Website, App, Social Media).
- **date**: Date the review was submitted.
- **reviewer_location**: City/location of the reviewer.
- **sentiment_score**: Computed polarity score (-1.0 to 1.0). Negative = negative sentiment, positive = positive sentiment.
- **reviewer_demographic**: Age group and gender combined (e.g., "26-35_Female").

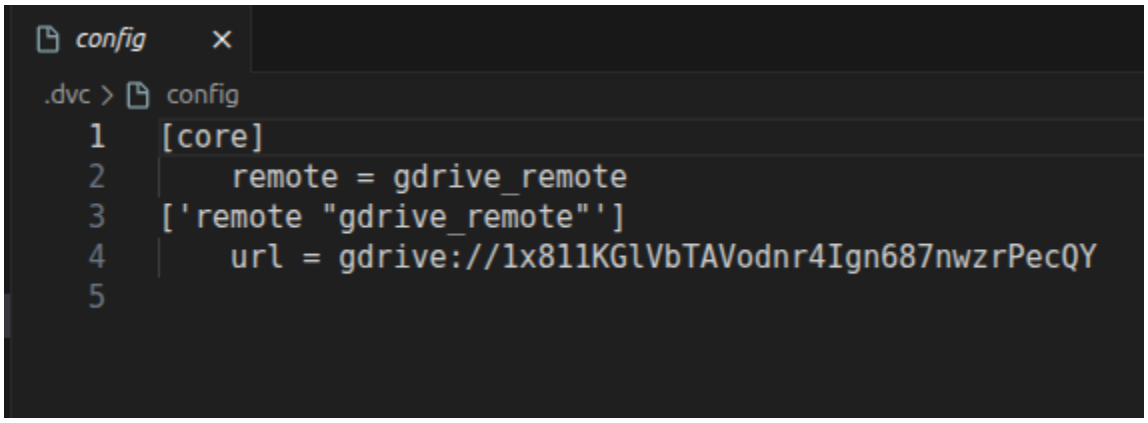
- **sentiment_category**: Sentiment label derived from sentiment_score (Positive, Negative, Neutral).
- **transaction_id**: Unique identifier for each purchase transaction (e.g., `TXN_00001`).
- **purchase_date**: Date and time of product purchase.
- **quantity**: Number of units purchased in the transaction.
- **price_per_unit**: Final price per unit after discounts.
- **region**: Purchase region (e.g., Chennai, Gurgaon).
- **customer_age_group**: Customer's age bracket (e.g., 18-25, 26-35).
- **customer_gender**: Customer gender (Male/Female).
- **season**: Season of purchase (e.g., Winter, Festive, Monsoon).
- **channel**: Sales channel (e.g., Online, App, Website Direct, Social Media).

product_id	product_name	brand	category	ingredients	clean_label	price	discount	units_sold	average_rating	num_re
TWT_021	71% Dark Chocolate Sweetened with Dates - Pack of 3	The Whole Truth	Dark Chocolate		Yes	1047	0	306	3.6	
TWT_017	No Added Sugar 5 Grain Muesli - Pack of - 350g x 1	The Whole Truth	Muesli	Seeds, Raisins, Oats, Dried Fruits, Almonds, Dates	No	349	0	1150	3.9	
TWT_024	Almond Millet Cocoa - Pack of 8 Millet Bars	The Whole Truth	Millet Bars	Cocoa Powder, Cocoa, Nuts, Dates, Millet	Yes	1000	0	478	4.2	
TWT_009	Hazelnut Cocoa Protein Bars - Box of 8	The Whole Truth	Protein Bars	Dates, Cocoa Powder, Cocoa, Whey Protein	No	1200	0	721	4.1	
TWT_024	Almond Millet Cocoa - Pack of 8 Millet Bars	The Whole Truth	Millet Bars		Yes	1000	0	1299	4.3	
TWT_024	Almond Millet Cocoa - Pack of 8 Millet Bars	The Whole Truth	Millet Bars		No	1000	0	1048	3.6	
TWT_021	71% Dark Chocolate Sweetened with Dates - Pack of 3	The Whole Truth	Dark Chocolate	Cocoa Butter, Cocoa, Almonds, Cocoa Powder	Yes	1047	0	696	3.7	
TWT_023	Double Cocoa Mini Protein Bars - Box of 12	The Whole Truth	Mini Bars		No	720	0	1226	3.6	
TWT_023	Double Cocoa Mini Protein Bars - Box of 12	The Whole Truth	Mini Bars		Yes	720	0	1465	4.4	
TWT_024	Almond Millet Cocoa - Pack of 8 Millet Bars	The Whole Truth	Millet Bars	Dates, Millet, Natural Flavoring	No	1000	0	870	4.2	
TWT_018	CRUNCHY - Peanut Spread With Dates - Box of 325g	The Whole Truth	Peanut Butter	Sea Salt, Peanuts, Dates	Yes	249	0	1402	3.6	
TWT_006	Pista Badaam 24 g Protein Powder - Pack of 1 KG	The Whole Truth	Protein Powder	Natural Flavoring, Whey Protein Isolate, Stevia, Lecithin	No	4499	9	1406	4.2	
TWT_021	71% Dark Chocolate Sweetened with Dates - Pack of 3	The Whole Truth	Dark Chocolate	Cocoa, Cocoa Butter, Almonds	Yes	1047	0	260	4.3	
TWT_020	CRUNCHY- Unsweetened Peanut Butter - Pack of 325g	The Whole Truth	Peanut Butter	Peanuts, Sea Salt, Dates	No	225	0	1023	4.1	
TWT_019	CREAMY- Unsweetened Peanut Butter - Pack of 325g	The Whole Truth	Peanut Butter	Sea Salt, Peanuts, Dates	Yes	225	0	1068	4	
TWT_024	Almond Millet Cocoa - Pack of 8 Millet Bars	The Whole Truth	Millet Bars		No	1000	0	1337	4.3	
TWT_018	CRUNCHY - Peanut Spread With Dates - Box of 325g	The Whole Truth	Peanut Butter		Yes	249	0	558	4.4	
TWT_016	Nuts, Fruits & Seeds Muesli - Pack of - 350g x 1	The Whole Truth	Muesli	Seeds, Raisins, Dried Fruits, Almonds, Dates, Oats	No	349	0	612	4	
TWT_019	CREAMY- Unsweetened Peanut Butter - Pack of 325g	The Whole Truth	Peanut Butter		Yes	225	0	840	4.1	
TWT_024	Almond Millet Cocoa - Pack of 8 Millet Bars	The Whole Truth	Millet Bars	Nuts, Natural Flavoring, Dates, Cocoa	No	1000	0	417	4.1	
TWT_015	Choco Fruit Crunch Muesli - Pack of - 350g x 1	The Whole Truth	Muesli	Oats, Raisins, Seeds, Almonds, Dried Fruits	Yes	499	0	900	3.9	
TWT_024	Almond Millet Cocoa - Pack of 8 Millet Bars	The Whole Truth	Millet Bars	Millet, Cocoa, Dates, Cocoa Powder, Natural Flavoring, Nuts	No	1000	0	736	4	
TWT_022	Almond Choco Fudge - Box of 10	The Whole Truth	Energy Bars	Cocoa, Dates, Nuts, Dried Fruits	Yes	750	0	296	3.9	

GSheet

DVC Configuration Files





Client ID and Client Secret kept in config.local to avoid credential leaks.

Risk Mitigation

Data Access Risks:

- **Incomplete Data**: Manual and extension-based scraping may miss fields.
  → *Use validation checks and spot verification.*
- **Website Changes**: Site structure updates can break scraping logic.
  → *Use flexible selectors and monitor regularly.*
- **Manual Effort Errors**: Human errors during scraping.
  → *Follow templates and checklists for consistency.*
- **Synthetic Data Limitations**: Faker data may lack realism.
  → *Use only where needed and label clearly.*
- **Tool Limitations**: Browser extensions may not handle all content.
  → *Use multiple passes and clean post-processing.*

Technical Risks:

- Large dataset storage → Use DVC with cloud storage
- Processing time → Implement parallel processing
- Reproducibility → Comprehensive logging and version control

Conclusion

Through this experiment, we established a comprehensive data science framework for analyzing The Whole Truth Foods' market position in India's clean-label food sector. We learned that consumer preference prediction requires multi-dimensional analysis combining social media sentiment, e-commerce behavior, and competitive data, rather than traditional single-source market research approaches. The integration of diverse data sources through DVC-versioned pipelines provides more accurate and cost-effective insights compared to existing solutions that rely on isolated platform analytics or expensive manual surveys. This experiment demonstrates how modern data science techniques can address real business challenges in product development, pricing optimization, and demand forecasting, establishing the foundation for data-driven decision making in the rapidly growing clean-label food market.