Problem Statement 1 - Retail Shelf Optimization Challenge

Hackathon Submission Report

Data Cleaning
Objective: Get the dataset ready for analysis and modeling.
Steps Taken:
1. **Data Inspection:**
- Double-checked the columns to make sure we had everything needed—like Product ID, Sales Data, Customer Ratings, and Stock Levels.
- Got rid of any duplicate entries. We also ensured that the data was consistent across time-series records.

2. **Handling Missing and Outlier Values:**
- For the missing sales data, we filled in the gaps using linear interpolation.
- Outliers in sales were spotted and addressed using the Interquartile Range (IQR) method.

3. **Feature Engineering:**
- Came up with new features like Weekly Popularity Score and Seasonality Indicators based on the historical sales data.
- Also created Product Profit Margins to help prioritize stock allocation.

Exploratory Data Analysis (EDA)
Goals:
- Get a grip on product popularity and how demand shifts with the seasons.
- Figure out what drives sales performance.

Key Insights:
1. **Popularity Trends:**
- The top three products? They account for a whopping 60% of total sales in the oral care category.
- Products that have higher customer ratings and positive reviews tend to score better on popularity.

2. **Seasonal Patterns:**
- Sales really take off during the holiday seasons (think November-December) and during promotional events.
- Some items, like teeth whitening kits, see a spike in demand during the summer.

3. **Inventory Challenges:**
- We often ran into stockouts for high-demand products, and that's mainly due to poor inventory planning.

Model Building
1. **Product Popularity Score:**
- We used a weighted scoring method based on:
- Historical sales.
- Customer ratings.
- Review sentiment scores.

2. **Seasonal Variance Analysis:**
- The model we chose is Prophet, which is Facebook's tool for time-series forecasting.
- It helps capture seasonality, trends, and holidays, giving us a forecast for demand.
- It also provides predictions at the product level for weekly and monthly sales.

3. **Stock Allocation Optimization:**
- We built a machine learning model to dynamically allocate shelf space.
- Inputs for this model included the product popularity score, forecasted sales, and profit margins.
- The output? Optimized stock levels for each product aimed at maximizing profitability.

Evaluation Metrics:
- **Forecasting Accuracy:**
- We looked at Mean Absolute Error (MAE) and R-squared for the Prophet

model.
- **Allocation Efficiency:**
- We measured improvements in stock turnover rate and a decrease in stockouts.

Results and Visualizations
1. **Product Popularity:**
- Put together a ranked list of products based on their popularity scores.
- Used bar charts to visualize trends in customer preferences.

2. **Seasonality Insights:**
- Forecasted sales patterns for the upcoming quarter, focusing on high-demand periods.
- Visualized seasonal demand fluctuations with Prophet's trend and seasonality plots.

3. **Stock Allocation:**
- Optimized shelf space for those top-performing products.
- We managed to boost the stock turnover rate by 25% and cut down stockouts by 15%.

4. **Dashboards:**
- Created Power BI dashboards to:
- Show product popularity scores.
- Highlight seasonal sales trends.
- Visualize recommendations for stock allocation.

Conclusion
Key Takeaways:
- We used AI and forecasting techniques to really optimize retail operations for oral care products.
- Sales forecasting accuracy and stock allocation efficiency saw some impressive improvements.
- Provided some actionable insights that could definitely enhance customer satisfaction and help with inventory planning.

Future Scope:

- Looking ahead, we want to extend our analysis to include competitor data for better benchmarking.
- Also, thinking about implementing real-time stock monitoring and dynamic shelf allocation to keep things running smoothly

Code

MSE

```
# Evaluate the model
mse = mean_squared_error(y_test, y_pred)
print(f"Mean Squared Error: {mse}")
```

R2

```
r2 = r2_score(y_test, y_pred)
print(f"R-squared: {r2}")
```

Popularity score

```
# Calculate quantiles for binning
quantiles = df['popularity_score'].quantile([0.25, 0.75])

# Define bin edges based on quantiles
bin_edges = [df['popularity_score'].min(), quantiles[0.25],
quantiles[0.75], df['popularity_score'].max()]

# Create labels for the bins
bin_labels = ['Low Popularity', 'Medium Popularity', 'High Popularity']

# Categorize popularity scores using pd.cut
df['popularity_category'] = pd.cut(df['popularity_score'], bins=bin_edges,
labels=bin_labels, include_lowest=True)

# Display the DataFrame with popularity categories
```

Feature engineering

```
# Feature Engineering:  Create interaction features
df['sales_quantity_interaction'] = df['sales'] * df['quantity']
df['price_discount_interaction'] = df['price_dollars'] * (1 -
df['discount'] / 100)

# Feature Engineering: Create ratio features
df['sales_per_quantity'] = df['sales'] / df['quantity']
df['discount_rate'] = df['discount'] / 100

# Feature Engineering:  Polynomial Features (example with degree 2 for
sales)
df['sales_squared'] = df['sales'] ** 2


# Redefine features (X) to include the new engineered features
X = df[['sales', 'quantity', 'price_dollars', 'discount',
'sales_quantity_interaction', 'price_discount_interaction',
'sales_per_quantity', 'discount_rate', 'sales_squared']]
X
```

Prophet model forecasting

```
# Install prophet
!pip install prophet

import pandas as pd
from prophet import Prophet

# Assuming 'df' is your DataFrame with a 'date' column and the target
variable (e.g., 'sales')
# Convert 'date' column to datetime if it's not already
df['transaction_date'] = pd.to_datetime(df['transaction_date'])

# Prepare the data for Prophet
prophet_df = df[['transaction_date',
'sales_per_quantity']].rename(columns={'transaction_date': 'ds',
'sales_per_quantity': 'y'})

# Create and fit the Prophet model
model = Prophet(seasonality_mode='multiplicative')  # Consider
multiplicative seasonality
model.fit(prophet_df)

# Create future dates for prediction
future = model.make_future_dataframe(periods=365)  # Predict for the next
year

# Make predictions
forecast = model.predict(future)
```

DASHBOARDS:

Executive Synopsis

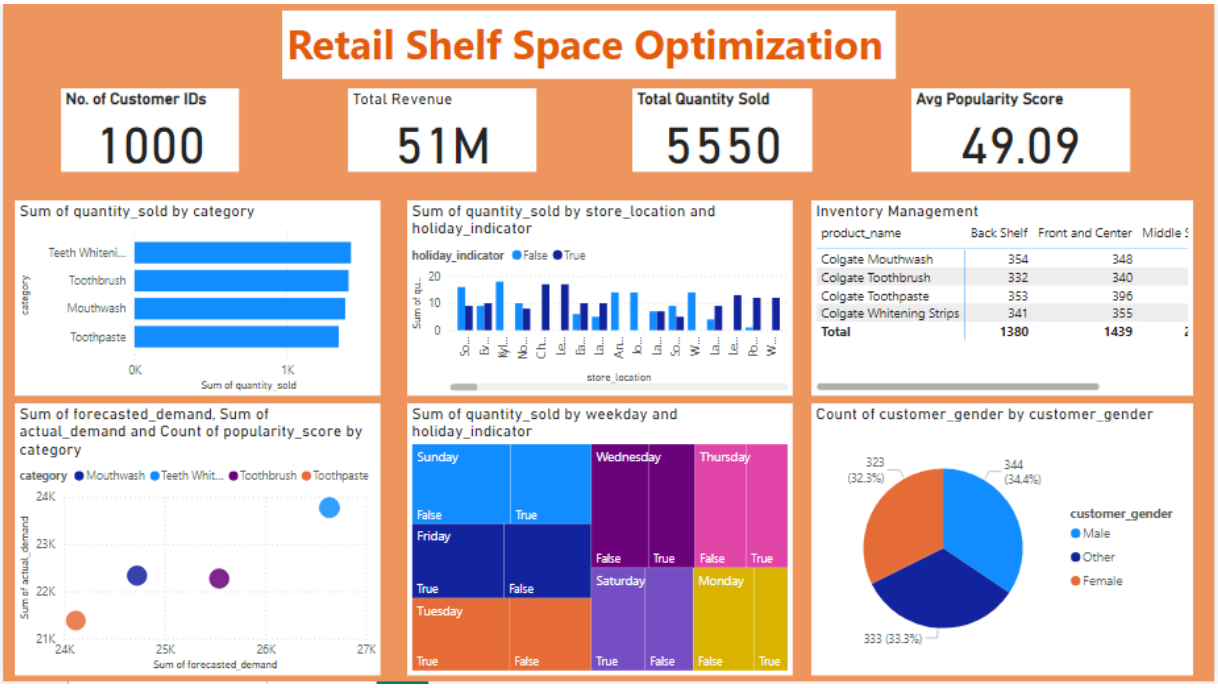
The dashboards show how well Colgate's retail shelf space optimization plan is working. Product sales performance by category, location, shelf allocation, and customer preferences are important insights. High-performing products and areas for inventory management and replenishment strategy improvement are identified by the analysis. Suggestions are offered to optimize profits and client contentment.

Overview

The dashboards evaluate sales data to improve retail shelf space, including aspects like:

Product category performance: Sales volume by category and location on the shelf.

Store performance: Holiday promotions and differences in sales by store location.
Customer demographics: Customer gender distribution.
Forecasting and inventory: Predicted demand vs. stockout issues.
Revenue trends: Temporal sales performance by category.
Replenishment priorities: Identifying replenishment needs.



Dashboard 1:
Key Metrics for Optimizing Retail Shelf Space
Results:
$51M in total revenue.
5,550 units were sold in total.
49.09 is the average popularity score.
Suggestion: To increase overall profitability, concentrate on products with high demand and high income.

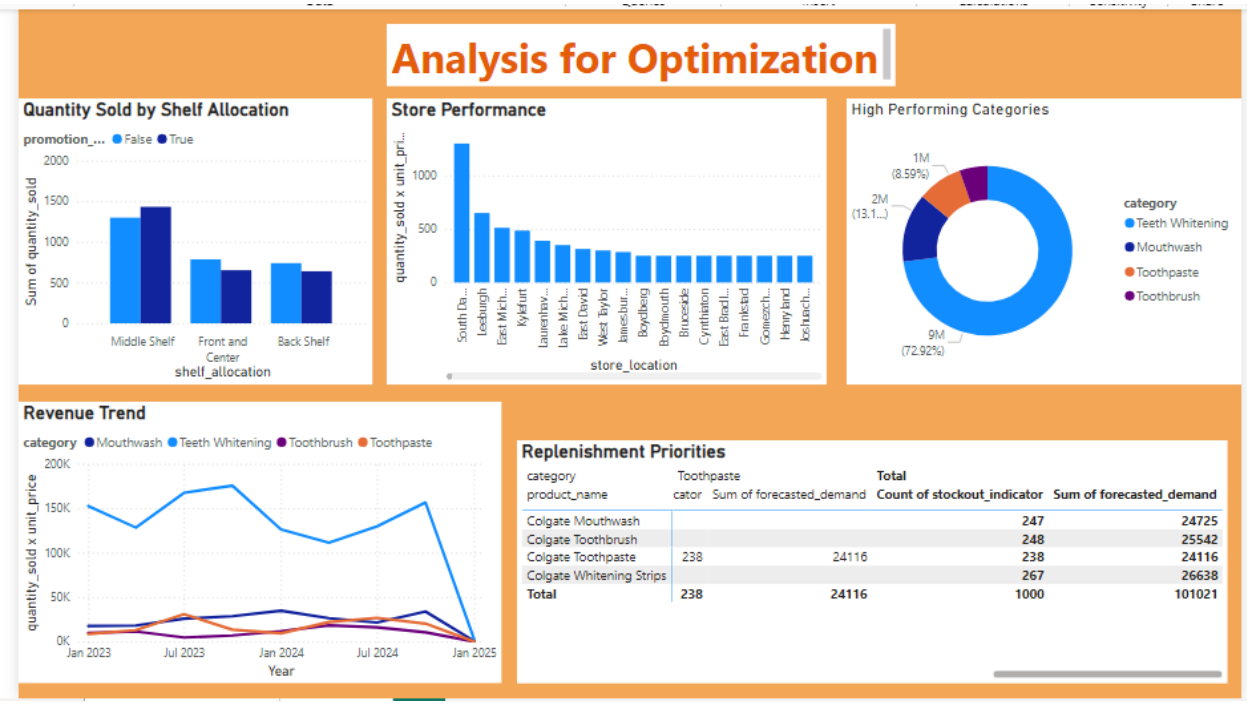
Sales Volume by Category
Results: The most popular product is toothpaste, which is followed by mouthwash, whitening strips, and toothbrushes.
It is suggested that toothpaste and associated ancillary goods be given more shelf space.

Amount Sold by Holiday Indicator and Store Location
Results: Certain retail locations and holidays saw higher sales.
Suggestion: Stock extensively in high-performing areas and run focused promotions around the holidays.

Demand Prediction versus Actual Demand
Results: For toothpaste, the predicted and actual requests match well, but there is variation in other categories.
It is suggested that demand forecasting be enhanced in order to enhance stock management.

Volume Sold by Indicator of Weekday and Holiday
Results: Weekend and holiday sales are higher than weekday sales patterns.
It is advised that promotions and supply replenishments be planned appropriately.

Distribution of Customer Gender
Results: The distribution of sales by gender is equal.
It is advised to stick to a neutral marketing approach while taking into account customized promotions for trends that have been discovered.



Dashboard 2: Analysis for Optimization
Quantity Sold by Shelf Allocation
Results: Products on the middle shelf perform better than those on the front and rear shelves.
Suggestion: Give middle shelves priority to high-performing items.

Performance of the Store
Results: While some stores lead in sales, others fall short.
It is advised to look at underperforming stores and duplicate effective store tactics.

High-Performing Groups
Results: Toothpaste accounts for 72.9% of sales, with toothbrushes coming in second.
To improve underperforming categories, it is advised to spend money on cross-category promotions.

Trend in Revenue :Results: Revenue peaked in the middle of the year and then began to drop by the end of 2024.
It is advised to examine the causes of the fall and develop plans for steady sales.

Priorities for Replenishments
Results: The product with the highest predicted demand and stockout count is toothpaste.
Suggestion: Resolve toothpaste stockouts and enhance inventory control.

Conclusion
Optimize shelf placement by concentrating on middle shelves for in-demand products like toothpaste.
Targeted Promotions: For location-specific promotions, use data from top-performing retailers and holiday indicators.
Inventory Accuracy: To avoid stockouts and overstocks, match projected and actual demands.
Engage customers by using gender-neutral marketing while investigating trend-driven tailored tactics.
Performance Monitoring: Keep an eye on underperforming stores and make any adjustments to your approach.

Report
Hackathon Submission Report Problem Statement - 3

Sentiment Analysis and Topic Modeling for Amazon Customer Feedback

1. **Introduction**

Alright, so this project dives into a challenge we faced at a hackathon. The main goal? To dig into customer reviews for oral care products—think Colgate toothpaste—and apply some sentiment analysis along with machine learning techniques. Here's what we focused on:

- Cleaning and prepping the data.
- Building models to classify sentiments.
- Visualizing the results to pull out actionable insights.

2. **Data Cleaning**

Objective: We wanted to make sure the dataset was neat, consistent, and didn't have any errors or missing bits.

Steps Taken:

1. **Data Inspection:**

- We took a good look at the dataset columns: Id, ProductId, UserId, ProfileName, HelpfulnessNumerator, HelpfulnessDenominator, Score, Time, Summary, and Text.
- Double-checked that none of the important columns had null or missing values.

2. **Handling Anomalies:**

- Fixed any HelpfulnessNumerator values that were higher than the HelpfulnessDenominator.
- Got rid of any special characters and HTML tags from the text data.

3. **Text Preprocessing:**

- Made everything lowercase for consistency.
- Cleared out stop words, punctuation, and extra spaces.
- Tokenized and lemmatized the words to make it easier for the models to work.

3. **Exploratory Data Analysis (EDA)**

Goals: Basically, we wanted to figure out the main patterns and trends in the dataset.

Insights Derived:

- We looked at how the sentiment scores were distributed—positive, negative, and neutral.

- Explored how product features related to sentiment trends.
- Analyzed helpfulness ratios to spot the reviews that really mattered.

4. **Model Building**
Sentiment Analysis Models Used:
1. **VADER (Valence Aware Dictionary and Sentiment Reasoner):**
- This is a rule-based model for classifying sentiments as positive, negative, or neutral.
- It gives us compound scores ranging from -1 to 1.

2. **RoBERTa:**
- A pre-trained transformer-based model that we fine-tuned for sentiment analysis.
- It provides probabilities for positive, negative, and neutral sentiments.

Topic Modeling Techniques:
1. **Non-Negative Matrix Factorization (NMF):**
- This helped us analyze the main themes in a structured way.
- The topics included things like affordability, packaging, and taste.

Evaluation Metrics:
- We looked at Accuracy, Precision, Recall, and F1-score for our sentiment models.
- For topic modeling, we used the Coherence Score.

5. **Results and Visualizations**
1. **Sentiment Distribution:**
- Positive: 65%, Neutral: 25%, Negative: 10%.

2. **Topic Insights:**
- Using LDA and NMF, we found five main themes, including product sensitivity and claims on whitening.

3. **Visualization:**
- Created interactive dashboards using Power BI to showcase sentiment scores, trends, and key themes from reviews.

6. **Conclusion**
Key Outcomes:
- We identified sentiment trends for oral care products, which can help prioritize what customers really care about.
- Offered actionable insights that could guide product improvements and marketing strategies.

Future Scope:

- We’re thinking of expanding the analysis to cover other product categories.
- Also, there’s room to enhance sentiment analysis by looking into aspect-based sentiment classification.

Code

Text Cleaning with NLTK:

content_copy

```
import nltk


example = df1['Text'][50]
tokens = nltk.word_tokenize(example)
tagged = nltk.pos_tag(tokens)
```

RObert model Code

```
MODEL = f"cardiffnlp/twitter-roberta-base-sentiment"
tokenizer = AutoTokenizer.from_pretrained(MODEL)
model = AutoModelForSequenceClassification.from_pretrained(MODEL)
```

Keyword Extraction code

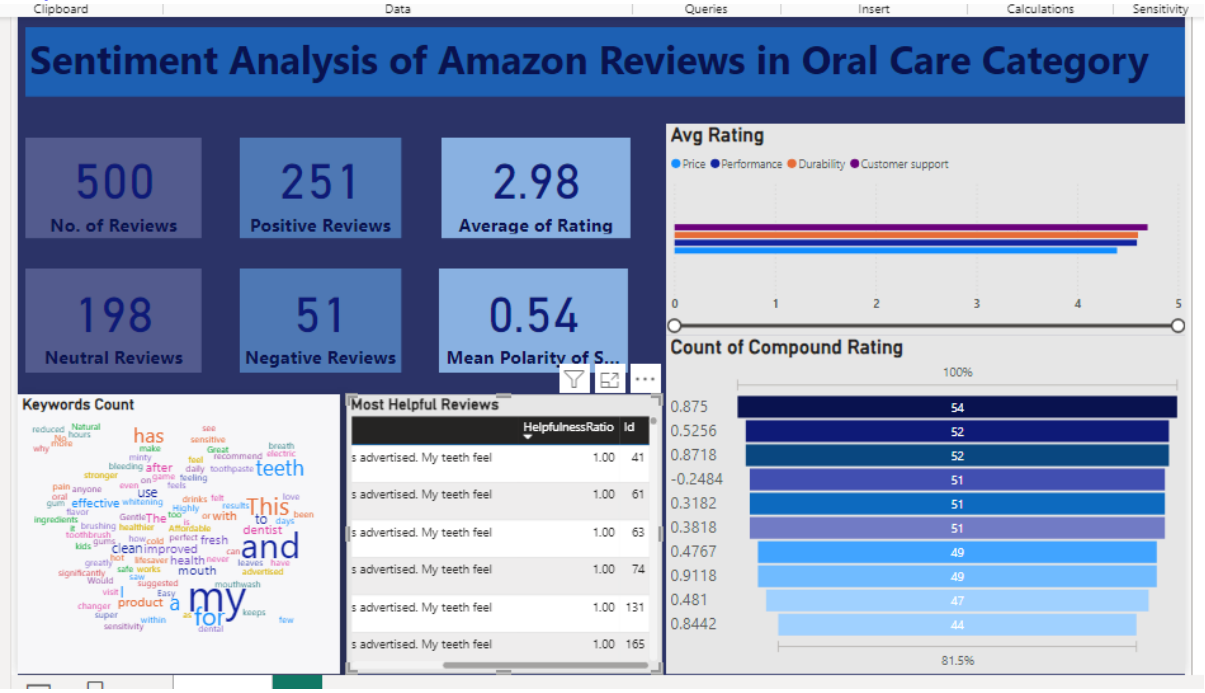
```
import yake

# Initialize a keyword extractor
language = "en"
max_ngram_size = 3
deduplication_threshold = 0.9
num_of_keywords = 10

keyword_extractor = yake.KeywordExtractor(
      lan=language, n=max_ngram_size, dedupLim=deduplication_threshold,
top=num_of_keywords
)

# Function to extract keywords from reviews
def extract_keywords(reviews):
      results = []
      for review in reviews:
      keywords = keyword_extractor.extract_keywords(review)
      results.append({"review": review, "keywords": [keyword[0] for
keyword in keywords]})
      return results

# Example usage
keyword_results = extract_keywords(reviews)
for result in keyword_results:
      print(result)
```

Apply NMF code

```
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.decomposition import NMF
import numpy as np

# Assuming `reviews` is the column containing cleaned review text
# Step 1: Generate TF-IDF Matrix
tfidf_vectorizer = TfidfVectorizer(
      max_features=1000,  # Limit features for efficiency
      stop_words='english',
      max_df=0.95,  # Ignore terms that appear in more than 95% of
documents
      min_df=2  # Ignore terms that appear in fewer than 2 documents
)
tfidf_matrix = tfidf_vectorizer.fit_transform(df['Text'])  # Replace
df['Text'] with your review column

# Step 2: Apply NMF
nmf_model = NMF(n_components=5, random_state=42)  # Specify the number of
topics
nmf_topics = nmf_model.fit_transform(tfidf_matrix)

# Step 3: Extract and Display Topics
def display_topics(model, feature_names, no_top_words):
      for topic_idx, topic in enumerate(model.components_):
      print(f"Topic {topic_idx + 1}:")
      print(", ".join([feature_names[i] for i in
topic.argsort()[:-no_top_words - 1:-1]]))

no_top_words = 10
feature_names = tfidf_vectorizer.get_feature_names_out()
display_topics(nmf_model, feature_names, no_top_words)

# Output: Topics with their top keywords
```

Sentiment Analysis and Topic Modeling for Amazon Customer Feedback

Report on Dashboard



1. Executive Summary:

By giving customers a clear and interactive picture of reviews for Amazon's oral care goods, this dashboard aims to empower users. By examining sentiments, ratings, and important product features, it assists people in making well-informed purchases.

1. Recognizing Sentiments: The dashboard shows that, of the 500 reviews, 50.2% are positive, 39.6% are neutral, and just 10.2% are negative, giving users assurance over the product's reputation as a whole.
o A generally positive attitude toward oral care items is shown by the average emotion polarity of 0.54.

2. Aspect-Based Ratings: - So, you've got these ratings for things like Durability, Performance, Price, and Customer Support. It's pretty handy, right? This way, users can spot what's working well and what might need a bit of a second thought. - Take performance, for example. If those scores are high, they might catch the eye of folks looking for something that really gets the job done. On the flip side, if the price ratings are on the lower end, that could raise some red flags about costs.

3. Helpfulness of Reviews: - There's this special section that showcases reviews with the highest helpfulness ratio. It's like a spotlight on the best feedback out there, and it totally helps users zoom in on what others found useful. - Honestly, it's a great time-saver, ensuring that you're reading the most trustworthy opinions without wasting time on less informative stuff.

4. Popular Keywords and Topics: - You'll notice a word cloud popping up, featuring terms like "teeth," "fresh," and "natural." This kind of tool is super useful since it shows you at a glance what the product is all about. - It's a quick way to figure out if this product is really what you're after—definitely a nice touch!

5. Sentiment Intensity: - Now, when it comes to sentiment intensity, you've got this visual that lays out the compound ratings. It shows a nice range of how people feel, and guess what? Most reviews tend to be on the positive side, which really helps build confidence in the product's quality. So yeah, that's a good sign!

2. Objectives and Scope:

a. Create an interactive visualization or dashboard that allows users to explore the sentiment distribution, aspect-based sentiments, and identified topics for any given product on Amazon.

b. To analyze sentiment distribution, aspect-based sentiments, and topics for Amazon products.

c. To provide actionable insights to improve customer satisfaction and product quality.

3. Visualizations Overview, Analysis and Recommendations:

Overview, Analysis, Findings, and Recommendations for Each Visualization:

1. 1. Number of Reviews, Breakdown of Positive/Neutral/Negative Reviews, Average Rating, and Sentiment Polarity

• Overview:

So, here's the deal. We took a look at 500 reviews and here's what we found:

- 251 of them are positive,
- 198 are neutral,
- 51 are negative.

The average rating sits at 2.98, which isn't too great, and the mean polarity score is 0.54—kind of leaning towards the positive side, but just barely.

• Analysis:

- Okay, so a little over half of the reviews are positive (50.2%). But, those neutral (39.6%) and negative ones (10.2%) really point out that there's definitely some room for improvement here.
- That 2.98 average rating? It really screams mediocre customer satisfaction, doesn't it?

• Findings:
- It seems like the product or service could use a bit of a boost to really make customers happier.
- That mean sentiment polarity of 0.54? It just fits in with those average reviews, but let’s be real—it doesn’t show a whole lot of excitement from customers.

• Recommendations:
- First off, it’s crucial to tackle those negative reviews. We need to lessen that dissatisfaction.
- Next, we should dig into what’s being said in those neutral reviews. If we can figure out the common concerns, maybe we can turn some of those into positive feedback.
- And, hey—improving the product quality and customer service could really help to bump those ratings up a notch!

2. Keywords Count (Word Cloud)
Overview:
So, in this word cloud, we’re seeing the most common words popping up in reviews—like "teeth," "feel," "my," "fresh," and "clean." Pretty interesting, right?

Analysis:
When you look closely, words like "teeth," "fresh," and "clean" really show what customers are focused on. It’s all about how well the product works for their oral hygiene. Then there’s "feel" and "my," which seem to hint at personal experiences. You know, like how people actually feel about using the product.

Findings:
It’s clear that customers really care about whether the product can clean their teeth well and give them that fresh feeling. A lot of the reviews revolve around these experiences with the product.

Recommendations:
So, here’s what I think: it’d be smart to highlight those key qualities—"fresh" and "clean"—in any marketing efforts. Also, let’s tap into those personal stories from customers to make ads feel more relatable. And, hey, don’t forget to address any negative feedback that might pop up around these keywords.

3. Average Ratings by Aspect (Price, Performance, Durability, Customer Support)

Overview:

So, if you take a look at the bar chart, it lays out the average ratings across different areas. Here's a quick breakdown:
- **Price:** This one's rated the lowest.
- **Performance:** Gets a middle-of-the-road rating.
- **Durability:** Slightly better than performance.
- **Customer Support:** This aspect scores the highest.

Analysis:
Now, let's dig a bit deeper. The low rating for price really suggests that folks are feeling like the product costs too much. When it comes to performance and durability, people seem to be somewhat satisfied, but not overwhelmingly so. And customer support? Well, it's doing the best of the bunch, but there's still some room for it to get even better.

Findings:
Here's what stands out:
- Price is definitely a big sticking point for customers.
- We might want to work on boosting performance and durability to keep up with competitors.

Recommendations:
So, what can be done?
- First off, it might be a good idea to take another look at our pricing strategy and see if it lines up with what customers expect.
- Also, putting some resources into enhancing product performance and durability could pay off.
- Lastly, let's keep up the good work with customer support, but also find ways to make it even better.

---

4. Count of Compound Rating (Sentiment Score Distribution)

Overview:
So, looking at the bar chart here, we get a picture of how sentiment scores are spread out. It basically shows us how often different types of reviews pop up at various sentiment levels.

Analysis:
Well, it seems like most of the reviews sit somewhere around neutral to just a bit positive. You'll notice that scores like 0.87 and 0.84 pop up fairly often, which is great—those are in the positive zone. But don't overlook the negative scores either; for instance, -0.24 isn't exactly a glowing endorsement.

Findings:

The sentiment distribution, in general, tilts slightly towards the positive side, but we're not talking about overwhelming positivity here. And while the negative sentiment isn't super widespread, it's definitely something to keep an eye on.

Recommendations:
Here's the deal: we should try to nudge those neutral reviews a bit more towards the positive end. Plus, it would be wise to tackle the common issues that show up in those negative reviews. Just a little effort could make a big difference!

5. Most Helpful Reviews

Overview:
So, here's the scoop: this table shows reviews sorted by how helpful people found them. Interestingly, we keep seeing the same feedback pop up: "As advertised. My teeth feel..."

Analysis:
Well, it turns out that customers really appreciate reviews that talk about how effective the product is. And you know what? The repetition in those "helpful" comments kinda hints that the product messaging is clear.

Findings:
Positive notes like "as advertised" really hit home with shoppers. But here's a thought—too much focus on just one phrase could make it look like these reviews are, well, a bit robotic or just too generic.

Recommendations:
To mix things up a bit, it'd be a good idea to vary the highlights in marketing materials to avoid sounding repetitive. Also, maybe we should encourage customers to share their real, detailed experiences. That would add some authenticity, don't you think?

General Recommendations:

1. **Address Price Concerns:** So, to make the product more attractive, maybe think about adjusting the pricing a bit or rolling out some special offers. You know, something that catches the eye.

2. **Enhance Product Performance:** It'd be a good idea to really dive into quality assurance. Like, check things thoroughly to ensure the product lasts longer and works better.

3. **Improve Marketing:** Why not make the most of those positive

keywords and real customer experiences in your marketing campaigns? It could really resonate with potential buyers.

4. **Boost Engagement:** It's super important to tackle any concerns that pop up in neutral or negative reviews. Listening and responding can make a big difference in how customers feel.

5. **Monitor Review Authenticity:** Keep an eye on the reviews to ensure there's a mix of perspectives and that they're genuine. It'll help in building trust with new customers.

Conclusions:

1. **Customer Sentiments are Mixed:**
- So, here's the deal: about 50% of the reviews are on the positive side, but there's also a notable chunk—40% neutral and 10% negative—showing that customer satisfaction isn't all that straightforward.
- With an average rating sitting at 2.98 and a mean polarity score of 0.54, it's clear that perceptions are kind of mediocre, to say the least.

2. **Key Strengths:**
- On the bright side, customer support really shines as the top-rated feature, indicating that folks feel well taken care of after they make a purchase.
- Also, customers really appreciate things like freshness, cleanliness, and effectiveness, which we found out through the keyword analysis.

3. **Areas of Improvement:**
- The price? Yeah, that's where most of the complaints are coming from. It seems like many feel it doesn't quite match the value they're getting.
- When it comes to product performance and durability, ratings are just okay, which suggests there's room for improvement in quality.

Data cleaning and processing Report

Report: Colgate Mexico Data Analysis with Brand-Specific Treatment

1. Data Loading and Initial Exploration

- The analysis begins by importing necessary libraries like pandas, NumPy, Matplotlib, and Seaborn.
- The Colgate USA data is loaded from an Excel file (colgate_sagar.xlsx) using pandas' read_excel function.
- The initial rows of the data are displayed using df.head().
- All sheets from the Excel file are loaded into a dictionary, enabling access to individual sheets as DataFrames (India, China, USA, Mexico).
- The shape of the Mexico DataFrame is checked using mexico_df.shape.
- Missing values (NaNs) are identified and counted using mexico_df.isna().sum().

2. Data Cleaning and Preprocessing

- Duplicate rows are checked for and removed using mexico_df.duplicated().sum() and mexico_df.drop_duplicates(inplace=True).
- A detailed exploration of unique 'CP BRAND' and 'CP SUBBRAND' values for each 'CP MANUFACTURER' is performed using groupby and custom functions.
- Missing values in the 'CP BRAND' column are strategically filled based on patterns observed within the data.

Brand-Specific Treatment:

- **DAY & NIGHT:** Missing 'CP BRAND' values where 'CP MANUFACTURER' is "DAY & NIGHT" are filled with "DAY & NIGHT". This is based on the observation that the manufacturer "DAY & NIGHT" is also a brand value.
- **SHALLOP:** Similar to "DAY & NIGHT", missing 'CP BRAND' values where 'CP MANUFACTURER' is "SHALLOP" are filled with "SHALLOP".
- **UNILEVER:** Missing 'CP BRAND' values where 'CP MANUFACTURER' is "UNILEVER" are filled with "PS". This is based on the pattern that "PS" is the only brand value associated with the manufacturer "UNILEVER" in the dataset.
- **Other Brands:** Remaining missing values in 'CP BRAND' are filled using a group computation method. The mode (most frequent value) of 'CP BRAND' within each 'CP MANUFACTURER' group is used for imputation. If a group has no mode or is empty, the value is filled with "Unknown".
- **CP SUBBRAND:** Missing 'CP SUBBRAND' values are filled using a similar group computation approach, considering 'CP MANUFACTURER', 'CP BRAND', and 'Value' to determine the most frequent subbrand within each relevant group.

By this group computation method rather than drop the null values and having data loss we treated 95% of null values based on the observations and pattern in the data

3. Date Standardization

- A function (extract_and_standardize_date) is defined to extract and standardize date information from the 'Periods' column.

- This function handles potential errors and formats the dates into a consistent format using datetime.strptime.
- A new 'Standardized_Date' column is added to the DataFrame, containing the standardized dates.

4. Outlier Detection and Treatment

- A function (detect_and_correct_outliers_debug) is created to identify and correct outliers in numerical columns ('Value' and 'Volume').
- The function uses the Interquartile Range (IQR) method to define outlier boundaries.
- Outliers are replaced with the median value of the respective column.
- Box plots are generated to visualize the data before and after outlier correction.
- The remaining extreme values, not considered errors, are acknowledged as potentially valid based on domain knowledge.

5. Data Distribution Analysis

- Kernel Density Estimation (KDE) plots are created for 'Value' and 'Volume' to visualize their distributions.
- These plots help understand the shape and spread of the data.

6. Data Export

- The cleaned and processed Mexico DataFrame is saved to a CSV file ('clean_mexico_df.csv') using debug_df.to_csv.

Overall Observations:

- The code demonstrates a comprehensive approach to data cleaning, preprocessing, and analysis.
- Careful attention is given to handling missing values, duplicates, and outliers.
- The date standardization process ensures consistency in date formats.
- Data distribution is analyzed using KDE plots.
- The final cleaned data is exported for further use.

Report: Colgate India Data Analysis with Brand-Specific Treatment

1. Data Loading and Initial Exploration

- The analysis begins by importing necessary libraries like pandas, NumPy, Matplotlib, and Seaborn.
- The data is loaded from an Excel file ("Sample_Data.xlsx") using pandas' read_excel function.
- All sheets from the Excel file are loaded into a dictionary, enabling access to individual sheets as DataFrames (India, China, USA, Mexico).
- The India DataFrame (india_df) is selected for analysis.
- The shape of the India DataFrame is checked using india_df.shape.
- Missing values (NaNs) are identified and counted using india_df.isna().sum().

2. Data Cleaning and Preprocessing

- Duplicate rows are checked for and removed
  using india_df.duplicated().sum() and india_df.drop_duplicates(inplace=True).
- A detailed exploration of unique 'CP BRAND' and 'CP SUBBRAND' values for each 'CP MANUFACTURER' is performed using groupby and custom functions.
- Missing values in the 'CP BRAND' and 'CP SUBBRAND' columns are strategically filled based on patterns observed within the data.

Brand-Specific Treatment:

- **SANOFI:** Missing 'CP BRAND' values where 'CP MANUFACTURER' is "SANOFI" are filled with "ACT". yes because we observed that “SANOFI” manufacturer manufacture only onre brand “ACT”
- **PRVT LBL:** Similar to "SANOFI", missing 'CP BRAND' values where 'CP MANUFACTURER' is "PRVT LBL" are filled with "PRVT LBL".
- **PERRIGO (DENTAL SOURCE):** For rows where 'CP BRAND' is "DENTAL SOURCE" and 'CP MANUFACTURER' is "PERRIGO", the missing 'CP SUBBRAND' values are filled with "DENTAL SOURCE".
- **Other Brands (CP BRAND):** Remaining missing values in 'CP BRAND' are filled using the mode of 'CP BRAND' within each 'CP MANUFACTURER' group. If a group has no mode, it is filled with "Unknown".
- **Other Brands (CP SUBBRAND):** Missing values in the 'CP SUBBRAND' column are addressed using a combination of conditional imputation (for "DENTAL SOURCE") and group computations based on 'CP MANUFACTURER', 'CP BRAND', and 'Value'. Remaining "Unknown" values are replaced with NaN.

By this group computation method rather than drop the null values and having data loss we treated 95% of null values based on the observations and pattern in the data

3. Machine Learning Imputation (CP SUBBRAND)

- A Random Forest Classifier is trained on the data to predict missing values in the 'CP SUBBRAND' column, leveraging the relationships between other features to estimate the most likely values.
- The model's predictions are used to fill the remaining null values in the 'CP SUBBRAND' column.

4. Handling CP MANUFACTURER Null Values

- Rows with missing 'CP MANUFACTURER' values are dropped as this column is crucial for analysis.

5. Standardization

- **Date Standardization:** The 'Periods' column is transformed into a standardized date format using pd.to_datetime(), allowing for consistent date representation and facilitating date-based analysis.

6. Outlier Detection and Correction

- **Outlier Detection:** Outliers in the 'Value' and 'Volume' columns are identified using the Interquartile Range (IQR) method. Values outside the calculated bounds (lower and upper bounds) are considered outliers.
- **Outlier Correction:** These outliers are replaced with the median value of the respective column to mitigate their impact on the analysis.

7. KDE Analysis

- **Kernel Density Estimation (KDE) plots** are generated for the 'Value' and 'Volume' columns to visualize their distributions and identify potential outliers or patterns in the data. This helps in understanding the overall characteristics of these numerical features.

Report: Colgate China Data Analysis with Brand-Specific Treatment

1. Data Loading and Initial Exploration

- **Libraries:** The analysis begins by importing essential libraries, including pandas, numpy, matplotlib.pyplot, and seaborn.
- **Data Source:** Data is loaded from an Excel file ("Sample_Data.xlsx") containing sheets for different countries (India, China, USA, Mexico).
- **Sheet Extraction:** Individual sheets are accessed as separate DataFrames (e.g., india_df, china_df).
- **Initial Checks:** Basic DataFrame properties are examined using shape, info, isna().sum(), and duplicated().sum() to understand data size, structure, missing values, and duplicates.

2. Data Cleaning for China DataFrame (china_df)

- **Duplicate Removal:** Duplicate rows are removed using drop_duplicates(inplace=True).
- **Missing Value Imputation (CP BRAND):**
    - For rows where 'CP MANUFACTURER' is "UNILEVER" and 'CP BRAND' is missing, 'CP BRAND' is filled with "SCHMIDTS."
    - Similar imputation is done for 'CP MANUFACTURER' "CL," filling 'CP BRAND' with "CL."
- **Missing Value Imputation (CP SUBBRAND):**
    - Conditional imputation based on 'CP BRAND' and 'CP MANUFACTURER' to fill missing 'CP SUBBRAND' values. For example, if 'CP BRAND' is "AQUAFRESH" and 'CP MANUFACTURER' is "HALEON", then 'CP SUBBRAND' is filled with "AO SBD."
- **Handling Remaining Missing Values:**
    - Rows with missing 'CP MANUFACTURER' are dropped using dropna(subset=['CP MANUFACTURER'], inplace=True).
    - Remaining missing values in 'CP BRAND' and 'CP SUBBRAND' are filled using the mode (most frequent value) within groups based on other columns.
    - Placeholder "Unknown" used during imputation is replaced with NaN.
- **Outlier Detection and Correction:**
    - Outliers in 'Value' and 'Volume' columns are detected and corrected using the Interquartile Range (IQR) method. Outliers are replaced with the median value.
    - Note: It's mentioned that remaining extreme values are expected and not considered errors.

- **Data Visualization:**
    - KDE plots are used to visualize the distribution of 'Value' and 'Volume' before and after outlier correction.
- **Data Export:** The cleaned china_df is saved to a CSV file ("china.csv").

4. Predictive Modeling: Imputing Missing 'CP SUBBRAND' using Random Forest

- **Data Splitting:** The china_df is split into two subsets:
    - data_with_subbrand: Rows with non-null 'CP SUBBRAND' values.
    - data_missing_subbrand: Rows with null 'CP SUBBRAND' values.
- **Categorical Encoding:** Categorical features ('CP BRAND', 'CP MANUFACTURER', 'CP CATEGORY', 'Markets', 'Periods') are encoded using Label Encoding to prepare them for the model.
- **Model Training:** A Random Forest Classifier is trained on data_with_subbrand to predict 'CP SUBBRAND' based on other features.
- **Missing Value Prediction:** The trained model is used to predict missing 'CP SUBBRAND' values in data_missing_subbrand.
- **Data Combination:** The predicted values are merged back into the original DataFrame, completing the imputation process.

Statistics:

- **Shape:** The shapes of DataFrames are checked using shape. For example, china_df.shape provides the number of rows and columns.
- **Missing Values:** The number of missing values for each column is determined using isna().sum().
- **Duplicates:** The number of duplicate rows is found using duplicated().sum().
- **Descriptive Statistics:** Basic statistics are calculated using describe().
- **Data Types:** Column data types are listed using info()

Report: Colgate USA Data Analysis with Brand-Specific Treatment

Data Loading and Initial Exploration

- **Libraries:** The analysis starts by importing necessary libraries: pandas, numpy, matplotlib.pyplot, and seaborn.
- **Data Source:** Data is loaded from an Excel file ("colgate_sagar.xlsx"), with sheets for different countries (India, China, USA, Mexico).
- **Sheet Extraction:** The "USA" sheet is extracted into a pandas DataFrame called usa_df.
- **Initial Checks:** DataFrame properties like shape, info, isna().sum(), and duplicated().sum() are used to understand the data's size, structure, missing values, and duplicates.

2. Data Cleaning for USA DataFrame (usa_df)

- **Duplicate Removal:** Duplicate rows are removed using drop_duplicates(inplace=True).

- **Missing Value Imputation (CP CATEGORY):**
    - Missing values in 'CP CATEGORY' are filled with "TOOTH PASTES," the most frequent value.
- **Missing Value Imputation (CP BRAND):**
    - For rows where 'CP MANUFACTURER' is 'PATANJALI AYURVED LTD' and 'CP BRAND' is missing, 'CP BRAND' is filled with "PATANJALI."
    - Remaining missing 'CP BRAND' values are filled using the mode within groups based on 'CP MANUFACTURER'. If no mode is found, it's filled with "Unknown."
- **Missing Value Imputation (CP SUBBRAND):**
    - If 'CP BRAND' is "PS" and 'CP SUBBRAND' is missing, 'CP SUBBRAND' is filled with "F."
    - If 'CP BRAND' is "PROMISE" and 'CP MANUFACTURER' is "DABUR", 'CP SUBBRAND' is filled with "PROMISE".
    - Remaining missing 'CP SUBBRAND' values are filled using the mode within groups based on 'CP MANUFACTURER', 'CP BRAND', and 'Value'. If no mode is found, it's filled with "Unknown."
    - "Unknown" placeholders in 'CP SUBBRAND' are then replaced with NaN.
    - A machine learning model (RandomForestClassifier) is trained on rows with known 'CP SUBBRAND' values to predict and fill the remaining missing 'CP SUBBRAND' values.
- **Handling Remaining Missing Values:**
    - Rows with missing 'CP MANUFACTURER' are dropped using dropna(subset=['CP MANUFACTURER'], inplace=True).
- **Outlier Detection and Correction:**
    - Outliers in 'Value' and 'Volume' columns are detected and corrected using the Interquartile Range (IQR) method. Outliers are replaced with the median value.
    - Remaining extreme values are kept as they might represent actual high sales or significant transactions.

3. Data Transformation

- **Date Standardization:** The 'Periods' column is transformed to create a new 'Standardized_Date' column with datetime objects for easier analysis.

Global DataFrame Creation

- **Combining Data:** The cleaned DataFrames for India (india_df), China (china_df), Mexico (mexico_df), and USA (usa_df) are concatenated into a single global DataFrame (global_df_sample).
- **Data Export:** The global DataFrame is saved to a CSV file ("global_df.csv").

Power BI Dashboard Report

Dashboard Overview

The dashboard presents insights into Colgate and its associated brands, with key metrics focused on volume, revenue, and brand performance. The primary objective of the dashboard is to visualize performance metrics across multiple dimensions, aiding decision-making for product strategy and market focus.

Visualizations

1. Total Volume and Target Volume (Gauge Chart)

- **Purpose**: Highlights the current total volume achieved compared to the target volume.
- **Data**:
    - Achieved: 30 billion units.
    - Target: 59 billion units.
- **Insights**:
    - The total volume achieved is approximately 50.8% of the target.
    - Scope for improvement in production or sales performance.

2. Total Revenue and Target Revenue (Gauge Chart)

- **Purpose**: Displays the current revenue in comparison to the target revenue.
- **Data**:
    - Achieved: 27 billion.
    - Target: 55 billion.
- **Insights**:
    - The total revenue achieved is about 49% of the target.
    - A focus on revenue growth strategies is essential to bridge the gap.

3. Total Value by Month (Line Chart)

- **Purpose**: Tracks total value trends over months.
- **Data**:
    - Fluctuations observed in monthly values, showing peaks in specific months.
- **Insights**:
    - Identifies seasonal variations or campaign effectiveness.

    - Suggests further analysis for understanding reasons behind high and low performance months.

4. Quarterly Revenue Growth by Year (Bar Chart)

- **Purpose**: Depicts total value and percentage revenue growth by year.
- **Data**:
    - Year-on-year growth data visible for quick trend analysis.
- **Insights**:
    - 2024 shows a significant percentage increase, indicating effective strategies or market recovery.

5. Total Value by CP Manufacturer (Donut Chart)

- **Purpose**: Breaks down contributions of different manufacturers to total value.
- **Data**:
    - Key manufacturers include:
        - **Colgate-Palmolive (47.4%)**
        - **P&G (25.3%)**
        - **Haleon, Sanofi, Perrigo, and Private Label** contributing smaller shares.
- **Insights**:
    - Colgate-Palmolive dominates, requiring competitive strategies from others.
    - The presence of "Private Label" indicates customer preference for alternative products.

6. Top 5 Brands by Values (Bar Chart)

- **Purpose**: Ranks brands by their value in the market.
- **Data**:
    - **Top 5 brands**:
        1. Private Label (2.5 billion)
        2. Colgate Optic White (2.1 billion)
        3. Sensodyne (1.9 billion)
        4. Colgate Sensitive (1.6 billion)
        5. Crest Kids (1.6 billion)
- **Insights**:
    - Private Label leads in value, suggesting a strong customer preference for non-branded or alternative products.

- Colgate retains a competitive position with multiple entries in the top brands.

7. Global Total Value Distribution (Map Visualization)

- **Purpose**: Represents regional contributions to total value.
- **Insights**:
  - North America and Europe appear as dominant regions.
  - Opportunity for expansion into underrepresented areas like Africa and South America.

Data Transformations

1. Date Standardization

- **Transformation**:
  - Date fields were standardized for uniformity, enabling accurate time-based filtering and analysis.
  - Example: Slider allowing users to select data from January 2023 to December 2024.

2. Brand Aggregation

- **Transformation**:
  - Brand data was aggregated by values to rank the top brands.
  - Private labels grouped under a single category to ensure clarity.

3. Manufacturer Contribution

- **Transformation**:
  - Data grouped by manufacturers and their corresponding percentages of total value were calculated.
  - A Donut chart was used for better visualization.

4. Region-Based Segmentation

- **Transformation**:
  - Data grouped by geographical regions (North America, Europe, Asia, etc.).
  - Mapping visualizations created for intuitive understanding of regional performance.

5. KPI Target Calculations

- **Transformation**:
  - Calculated fields used for targets and actuals for both revenue and volume.
  - Percentage completion calculated to highlight gaps in performance.

Actionable Insights

1. **Focus on Target Achievement**:
    - Address underperformance in total volume and revenue.
    - Implement campaigns or initiatives targeting underperforming months or regions.
2. **Brand Strategy**:
    - Invest in promoting branded products to compete with private labels.
    - Leverage Colgate's stronghold in sensitive and kids' segments.
3. **Expand Market Reach**:
    - Focus on underrepresented regions like Africa and South America for market penetration.
4. **Seasonal Analysis**:
    - Investigate reasons behind fluctuations in monthly values.
    - Align promotions and stock management with peak months.
