

Rebel Sport (Rebel) is a leading sporting goods retailer operating in Australia and New Zealand, specializing in a diverse range of sports equipment, clothing, footwear, and accessories. Products are designed to cater to both amateur and professional athletes, as well as individuals with active lifestyles. The company offers items tailored to various sports, including football, cricket, tennis, basketball, swimming, fitness, camping, and outdoor activities. Its primary activities include retail sales, online shopping, an extensive product range, promotional events, and the provision of expert advice and assistance. Since its

establishment in 1985, Rebel has expanded significantly, becoming Australia's largest retailer of sporting goods. It ranked among the top three sports retailers in Australia as of February 2025.

Social media analytics (SMA) has been leveraged by Rebel to enhance operational efficiency, generate data-driven insights, and improve customer engagement, thereby boosting sales and strengthening brand loyalty. Specifically, SMA benefits include efficient inventory management, which enables the monitoring of stock levels, tracking of product movement, and automation of reordering processes. Sales systems are streamlined to process payments rapidly, maintain accurate sales records, and provide real-time sales performance data. Customer relationship management is strengthened by capturing customer data, purchase history, and preferences, allowing personalized marketing, targeted promotions, and improved customer satisfaction. Enhanced data analytics supports sales trend analysis and provides insights into customer behavior and inventory management. The company's online presence has also been improved, enabling the effective management of its website, online store, and digital marketing initiatives. Supply chain management has benefited from better coordination with suppliers, shipment tracking, logistics management, and timely deliveries. Furthermore, SMA supports informed decision-making by providing access to performance metrics, financial data, market trends, and customer behavior.

Rebel's strategic goals include **"Grow the Core Brand"** and **"Leverage Closeness to Our Customer"** ( [Annual report](#), 2024). Both objectives emphasize deepening customer relationships through data-driven personalization and enhancing brand relevance. By integrating SMA insights with historical customer data, Rebel tailors product recommendations, highlights trending items, and offers exclusive member discounts. Customer interactions on platforms such as YouTube, Facebook, Reddit, and Instagram, as well as reviews posted on Rebel's website, are analyzed to track sentiment and engagement. These interactions are monitored through metrics such as likes, shares, comments, and campaign hashtags, including *#FootballsEverything* and *#SportsCalling*, to optimize marketing efforts and foster stronger connections with the customer base.

## 1.2 Explanation of Business Impact

Text analysis has been employed by Rebel Sport to mine valuable insights from unstructured data, providing measurable business benefits by enhancing marketing effectiveness, optimizing inventory, and improving customer satisfaction.

### 1. Increased Sales and Marketing ROI:

By tracking customer engagement across social media channels, Rebel gains real-time insights into current trends. This enables the company to tailor its promotions, product recommendations, and marketing campaigns to align with

shifting customer preferences. For example, if SMA tools detect a growing interest in football gear due to an upcoming tournament, Rebel can adjust its promotional content and optimize inventory to capitalize on this demand. This strategic alignment improves sales, maximizes ROI on marketing campaigns, and increases average basket sizes.

## **2. Enhanced Customer Experience and Loyalty:**

Rebel's use of sentiment analysis helps identify and address customer concerns, enhancing the overall shopping experience. Tracking sentiment also allows Rebel to proactively respond to positive or negative shifts in customer preferences, strengthening brand loyalty. Personalized marketing—such as exclusive member discounts based on past purchase history—further enhances customer retention by making each interaction more relevant and engaging.

## **3. Campaign Effectiveness Evaluation:**

SMA allows Rebel to assess the performance of its marketing campaigns by tracking sentiment, engagement, and customer discussions. By analyzing which campaign messages resonate most with audiences, Rebel can refine its future campaigns, optimizing both messaging and targeting to drive higher engagement and conversion rates.

## **4. Improved SEO and Brand Visibility:**

In addition to campaign evaluation, Rebel uses text analysis tools to extract key insights from unstructured data, such as popular search terms, trending hashtags, and customer reviews. By incorporating these insights into its SEO strategy, Rebel can improve website content, enhance search engine rankings, drive organic traffic, and boost brand visibility.

# **1.3 Use of Real Data and Specific Example**

A prime example of Rebel Sport's effective use of social media analytics is its \*Sport Is Calling\* campaign, launched in 2020 and reiterated in 2024. This campaign featured inspiring stories of real-life Australian athletes and adventurers, including Brendan Cullen and Olympian Sinead Diver. Its emotional storytelling and strategic timing ahead of the Paris Olympics created substantial social media buzz, reinforcing Rebel Sport's position as a champion of Australian sports.

The campaign's impact was evident in the numbers:

- Over **10,900 Instagram posts** used the hashtag *#SportsCalling*.
- More than **100 YouTube videos** were created by users and influencers.
- The top YouTube videos garnered over **1,100,000 views**, significantly boosting Rebel's online visibility and engagement ([Rebel Sport](#), 2024).

By tracking these engagement metrics, Rebel gained actionable insights into which campaign stories resonated most with their audience. This data allowed Rebel to refine its content strategy, optimize future campaigns, and enhance its reputation as a supporter of Australian athletes. The campaign's success not only increased brand visibility but also strengthened customer connections, contributing to higher sales and long-term brand loyalty.

### **Conclusion: SMA as a Strategic Asset**

Rebel Sport's strategic use of social media analytics and historical data demonstrates how data-driven insights can enhance business performance. By integrating sentiment analysis, trend tracking, and real-time engagement metrics into their decision-making processes, Rebel improves marketing effectiveness, optimizes inventory, and personalizes the customer experience. This approach positions Rebel for continued growth and ensures that it remains at the forefront of the competitive sports retail market.

## **2. App Feature Proposal: Rebel GearMatch**

### **2.1 Novelty and Creativity: The Power of Comparison and Price Matching**



Photo Source: [Comparison example](#), image from <https://runtothefinish.com>

In today's competitive retail environment, customers are increasingly price-conscious, with a significant portion of their shopping experience revolving around finding the best deal. According to [Koen van Gelder](#), an conducted survey across the United States, United Kingdom, France, Germany, and Australia showed that, at least 80% of online shoppers compare prices before making a purchase, demonstrating the strong demand for a seamless comparison tool ([Koen van Gelder](#), 2024).

However, the process of comparing prices across multiple platforms and websites can be time-consuming and frustrating. Customers often have to visit several websites, social media platforms, and read numerous reviews before making a final decision. Rebel Sport can address this challenge by introducing **Rebel GearMatch**, a powerful price comparison tool

designed to ensure customers always get the best value without ever leaving the Rebel Sport website.

## Rebel GearMatch: A New Tool for Price Comparison and Product Selection

**Rebel GearMatch** is an innovative product comparison tool designed to simplify the purchasing decision process. This tool not only compares product features but also matches or beats competitors' prices, ensuring Rebel Sport customers get the best deals available in the market.

### Value to the Customer

To determine what's valuable, we use a simple formula:

$$Value^* = \frac{\text{Relevance} + \text{Timeliness}}{\text{Loss of privacy}} \text{Trust}$$

\* to the customer

The value a customer perceives in a product depends on its relevance and timeliness, relative to the personal effort or data they need to share to get this information. Essentially, customers view value as the balance between how relevant and timely the product information is versus the "cost"—that is, the effort or personal data required to obtain it.

**Rebel GearMatch** Rebel GearMatch would be integrated directly into the Rebel Sport website, offering customers a seamless experience without needing to navigate to other platforms. The tool goes beyond comparing basic product features by also providing real-time competitive pricing data. Whether a customer is searching for running shoes, football gear, or fitness accessories, Rebel GearMatch will present side-by-side comparisons of up to three products from Rebel Sport.

Key features of **Rebel GearMatch** include:

- **Price Comparison:** Real-time pricing information across products.
- **Product Features:** Side-by-side comparison of specifications, customer ratings, and reviews.
- **Promotions:** Highlight current discounts or exclusive deals available through Rebel.
- **Product Recommendations:** Based on previous purchases or browsing behavior, suggesting alternatives with better value for money.

By offering price transparency and the opportunity for price matching, Rebel GearMatch ensures that customers can make informed decisions without feeling the need to search across multiple sites. This ultimately leads to a smoother purchasing experience and an increased likelihood of conversion.

## 2.2 Clarity and Detail

The primary goal of Rebel GearMatch is to give customers all the tools they need to make an informed, confident purchase without leaving the Rebel Sport website. The app would allow customers to compare up to three similar products from different brands or price points and present a comprehensive breakdown of each product's features, reviews, and pricing. This makes it easier for customers to evaluate what's best for their needs and preferences. For example, if you are looking for a running shoes of a particular brand and you want the best deal available on that shoes. Rebel will base on historical data to automatically highlight the price or review stars or the product characteristics, make it easier decision, encouraging the customer to complete the purchase on the site.

Rebel GearMatch simplifies the shopping experience by allowing customers to compare up to three products directly on the website, including key details like price, reviews, and specifications.

The comparison tool will use data such as previous purchase behavior, product interests, and social media interactions to offer personalized comparisons. This could drive higher engagement, as users will see products that align with their preferences. Research by McKinsey & Company (2021) indicates that personalized experiences can increase revenue by up to 15%, showing the potential for improved sales.

By integrating upselling and cross-selling features, Rebel GearMatch can also encourage customers to explore higher-value products or complementary items, increasing the average order value.

### **User Experience Benefits:**

By bringing all relevant product details into one easy-to-use comparison tool, Rebel GearMatch eliminates the need for customers to toggle between multiple tabs and websites. This streamlined, transparent comparison process makes it faster and easier for customers to choose the right products, reducing decision fatigue and cart abandonment.

## **2.3 Alignment with Business Objectives**

As data-driven decision-making becomes more prevalent, online shoppers increasingly seek detailed and accessible product information before making a purchase. Understanding this trend, Rebel Sport can meet customer needs by providing clear, customer-focused product comparisons, which can increase conversion rates and enhance customer loyalty.

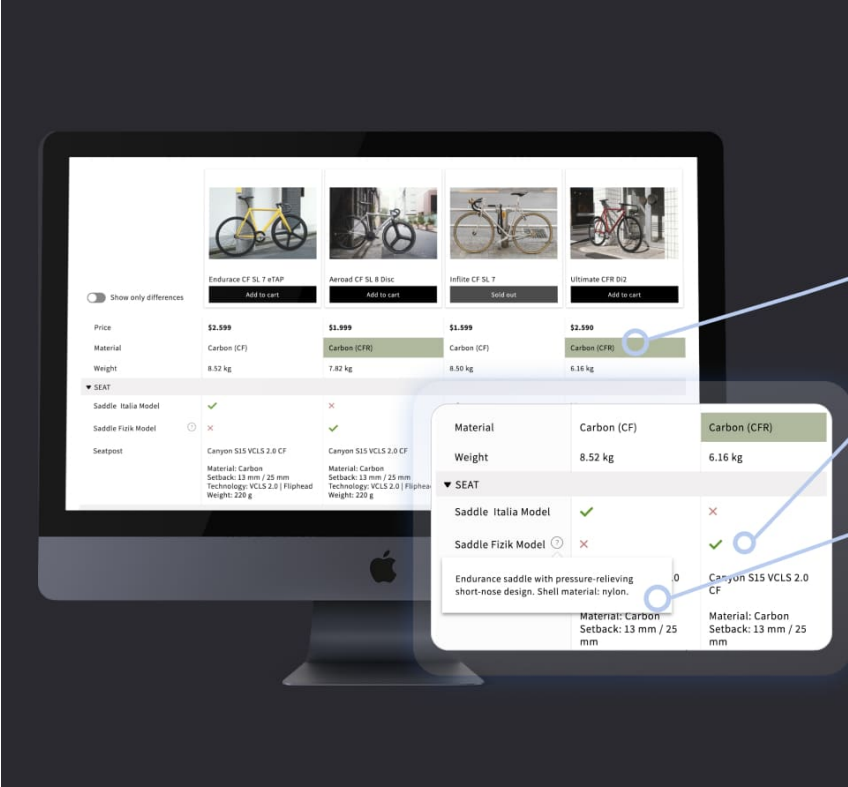
Rebel GearMatch supports Rebel Sport's business objectives by aligning with key strategies like **"Grow the Core Brands"** and **"Leverage Closeness to Our Customer."**

- **Increasing Customer Loyalty and Satisfaction:** By providing a seamless price comparison experience and offering price matching, Rebel Sport enhances customer trust. Customers will feel assured they are getting the best deal, which increases the



likelihood of repeat purchases and long-term loyalty.

- **Boosting Conversion Rates:** With price transparency and the ease of comparison, Rebel GearMatch will reduce cart abandonment and hesitation, leading to higher conversion rates. Customers will appreciate the effort Rebel Sport puts into ensuring the best value, leading them to complete their purchase with confidence.
- **Competitive Advantage:** Offering price matching within a streamlined comparison tool helps Rebel Sport differentiate itself from competitors. It ensures customers do not have to shop around, giving Rebel Sport a competitive edge in customer service and pricing.
- **Improving Sales Performance:** Through upselling and cross-selling features, Rebel GearMatch can encourage customers to consider higher-value products or complementary items, increasing the average order value (AOV). Research has shown that personalized experiences like these can drive significant revenue growth.



## BOOST SALES

Promote high-margin products

Emphasize the unique strengths

Help customers make informed decisions

Photo Source: [App for Comparison](https://apps.shopify.com/comparable), compare by custom option, image from <https://apps.shopify.com/comparable>

### 3. Rebel GearMatch: An Innovative and useful App feature

Rebel GearMatch is an innovative and highly useful app feature designed to enhance the shopping experience by leveraging advanced technologies. By incorporating Artificial



Intelligence (AI) and Machine Learning (ML), the tool transforms how customers compare products, creating a seamless, personalized, and efficient shopping process.

### 3.1 Explanation of Innovation

Rebel GearMatch distinguishes itself through the **innovative application of AI and ML technologies** to deliver personalized, dynamic product comparisons. Unlike traditional comparison tools, which typically offer static, generalized options, Rebel GearMatch tailors product recommendations and comparisons based on individual user behavior and preferences. This personalization creates a more intuitive and relevant experience for each customer.

Additionally, Rebel GearMatch functions as a centralized platform by integrating multiple data points—such as *price, promotions, features, specifications, and customer reviews* into a single, user-friendly interface. By consolidating this information, customers can access a holistic view of the products they are considering. This reduces the need to navigate across various websites, saving time and effort while enhancing convenience.

### 3.2 Discussion of Usefulness

The usefulness of Rebel GearMatch is reflected in its ability to simplify and streamline the shopping process for customers. The app enhances the shopping journey by offering several key benefits:

- **Centralized Comparison:** Customers can compare products side-by-side, from pricing and features to customer reviews, all in one place. This comprehensive approach allows users to make informed decisions more quickly and confidently.
- **Interactive, User-Friendly Interface:** Designed for simplicity, the interface is intuitive, enabling customers to explore comparisons, discover new products, and make purchases without feeling overwhelmed. By reducing complexity, the tool helps users make faster purchasing decisions without switching between multiple websites or tabs.
- **Personalized Recommendations:** AI and ML technologies analyze customers' past browsing and purchasing behavior to generate tailored product suggestions. This personalized approach increases the likelihood that users will find items that align with their preferences, thus enhancing satisfaction.

### 3.3 Supporting Evidence and Examples

The effectiveness and potential impact of Rebel GearMatch are supported by empirical studies and real-world evidence:

- **Increased Revenue Through Personalization:** Research by McKinsey & Company (2021) shows that personalization can lead to a 5-15% increase in revenue, proving that

tailoring product recommendations enhances sales performance.

- **Consumer Preferences for Relevant Recommendations:** According to Accenture, 91% of consumers are more likely to shop with brands that offer personalized recommendations and relevant offers, which is exactly what Rebel GearMatch delivers ([Jose, H.](#), 2023).
- **Increased Customer Retention Equals Higher Profits:** A study by Bain & Company indicated that a 5% increase in customer retention could lead to a 25-95% increase in profits ([Bain & Company](#), 2024). By delivering a personalized, streamlined shopping experience, Rebel GearMatch can enhance customer loyalty, contributing to long-term business profitability.

Through its innovative design, practical benefits, and evidence-backed outcomes, Rebel GearMatch provides a valuable tool that strengthens Rebel Sport's competitive position while fostering customer engagement and loyalty.

## 4. Challenges of Implementing the Proposed Application

### 4.1 Identification of Challenge: Data Privacy and Security

The integration of personalized features, such as product comparisons based on past purchases and browsing behavior, brings with it concerns regarding data privacy. Rebel GearMatch would require the collection of significant amounts of personal data to deliver tailored recommendations. This data collection raises privacy concerns, especially as consumers become more cautious about their personal information.

### 4.2 Depth of Discussion on Challenge

Ensuring compliance with data protection regulations such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) is complex. These regulations demand that businesses obtain explicit consent from users before collecting their data and offer the ability to opt-out. For instance, Rebel Sport must inform customers of the types of data collected (e.g., browsing history, purchase history) and how it will be used, including offering clear options for customers to manage or delete their data. Failure to do so could lead to legal repercussions, customer distrust, and damage to Rebel Sport's reputation. Additionally, safeguarding this data from breaches or misuse is crucial.

### 4.3 Suggestions for Mitigation

To mitigate privacy concerns, Rebel Sport should implement robust data security protocols, including encryption and secure data storage. It's important to obtain explicit user consent before collecting data and ensure transparency around data usage through clear privacy policies. Users should be given control over their data, with the option to manage or delete it as needed. Regular compliance checks and audits should also be carried out to ensure ongoing adherence to data protection regulations, maintaining customer trust.

## 5.A Objectives of using Reddit data

In today's competitive retail landscape, gaining a deeper understanding of customer sentiment, preferences, and emerging trends is essential for enhancing customer loyalty. Rebel Sport seeks to achieve this by leveraging data from real-time, user-generated sources.

Reddit, a platform with over **365 million weekly active users** ([Backlinko Team](#), 2025), provides an extensive pool of real-time data. Insights can be extracted from relevant subreddits such as **r/sports**, **r/Fitness**, and **r/RunningShoeGeeks**. These communities offer valuable discussions on sports enthusiasts' preferences, feedback on specific brands and

products, and trends in fitness and athletic gear.

By analyzing these conversations, insights into customer opinions, product satisfaction, and emerging fitness interests can be gathered, allowing Rebel to tailor its offerings, enhance engagement, and improve customer satisfaction.

## Objective: Identify Emerging Trends and Product Demand

### Clarity of Objective:

The objective of using Reddit data is to monitor relevant discussions in fitness- and sports-related subreddits (*r/Fitness*, *r/RunningShoeGeeks*, and *r/sports*) to identify products, brands, or sports activities that are gaining popularity. By tracking customer conversations in real time, Rebel Sport aims to detect shifting preferences, emerging trends, and rising product demand early in the trend cycle.

### Relevance of Objective:

This objective is directly relevant to Rebel Sport's context as a leading sporting goods retailer. Staying informed about current trends enables the company to optimize its product offerings, ensuring that trending products are stocked, marketed, and promoted at the right time. This alignment enhances customer engagement, satisfaction, and brand relevance by offering products that reflect evolving market demands.

### Justification of Objective:

Justifying this objective is critical to Rebel Sport's broader strategic goals. Reddit, as a platform with over 365 million weekly active users, provides real-time, user-generated data that offers authentic insights into customer preferences ([Backlinko Team](#), 2025). By leveraging these discussions, Rebel Sport can mitigate inventory risks, reduce excess stock, and maximize sales by stocking products that resonate with customer demand.

Moreover, early identification of emerging trends provides Rebel Sport with a competitive edge by enabling it to adapt marketing campaigns swiftly, highlight relevant products, and position itself as a trend leader. This proactive approach aligns with Rebel Sport's strategic drivers to "*Grow the Core Brands*" and "*Leverage Closeness to Our Customer*" ([Annual Report](#), 2024), ultimately boosting customer loyalty, sales, and profitability.

## ✓ 5. B Data Required for Customer Insights and Loyalty Analysis

### 5. B.1 Identification of Relevant Data

Rebel Sports Australia is sports brand retailers that category product into footwear, fitness and brands like Nike, Addidas, Asics, Puma, Under Armour and New Balance. To gather insights on customer sentiment and feedback about Rebel Sport's products and relevant trends, these subreddit discussions are the best and suitable.

## Subreddits for Targeted Insights

1. r/sports - 22M+ members, Sports News and Highlights from the NFL, NBA, NHL, MLB, MLS, and leagues around the world.
2. r/Fitness – 12M+ members. Contains product reviews, workout advice, and fitness trends.
3. r/RunningShoeGeeks – A smaller but highly targeted subreddit discussing running gear, shoes and accessories.

These subreddits are rich sources of information, including user-generated product reviews, price comparisons, and brand loyalty discussions.

## Data Fields to Be Collected

To maximize the value of Reddit data, the following fields will be extracted from posts and comments:

### 1. Post data

- Title: Useful for identifying trending topics.
- Post body (text): Contains valuable customer feedback, product opinions, and suggestions.
- Subreddit name: To differentiate discussions by community.
- Number of upvotes (score): Indicates engagement and community interest.
- Number of comments: Helps identify posts with viral potential.
- Created date: Useful for time-series analysis.
- URL: To track original posts.

### 2. Comment Data:

- Comment text: Crucial for understanding customer sentiment and engagement.
- Comment score (likes): Shows which comments resonate most with the audience.

3. **Product and Brand Mentions** To link Reddit insights with Rebel Sport's offerings, identify product- and brand-specific mentions:

- **Brands:** Nike, Adidas, Puma, Asics, New Balance, Under Armour, etc.
- **Products:** Running shoes, sneakers, fitness gear, football shoes, sportswear.
- **Keywords:** Relevant keywords relate to Rebel product (e.g., "running," "training shoes").

## 5. B.2 Data Acquisition Process

The data can be collected using web scraping and Reddit APIs, with the following process:

### 1. Using Reddit's Official API:

- **Access:** Apply for API credentials through Reddit's developer portal.
- **Endpoint Usage:** Use endpoints like `/r/{subreddit}/comments` to collect real-time data from relevant subreddits.
- **API Parameters:** Set parameters to filter data by engagement (e.g., top posts).

### 2. Web Scraping:

- Tools like PRAW (Python Reddit API Wrapper) will be used to collect historical data if necessary.
- Implement scraping scripts with proper respect for Reddit's terms of service, avoiding rate-limit violations.

### 3. Data Cleaning and Preprocessing:

- **Remove Spam/Low-Value Data:** Filter out posts/comments with low scores or flagged as spam.
- **Standardize Text Data:** Convert all text to lowercase, remove special characters, and tokenize for analysis.
- **Identify Product Mentions:** Use Named Entity Recognition (NER) and keyword matching to extract brand and product names.

```
# Install required packages:  
!pip install asyncpraw pandas nest_asyncio
```

 [Show hidden output](#)

```
# Import required libraries:  
import asyncio
```

```

import asyncpraw
import pandas as pd
import nest_asyncio
from datetime import datetime
import re

# Allow nested event loops in Colab:
nest_asyncio.apply()

# Reddit API credentials:
client_id = '_mEvCBQLeRsUrN5nmdcD6A'
client_secret = 'ezp0D4hfMP0_jD2Sdd7xX16T_YKXvQ'
user_agent = 'Confident-Ad8604'

# Define subreddit and the number of posts to fetch:
subreddits = ["sports", "Fitness", "RunningShoeGeeks"]
num_posts = 300

# Create an asynchronous function to fetch posts and comments:
async def fetch_posts_and_comments(reddit, subreddit_name):
    subreddit = await reddit.subreddit(subreddit_name) # Access subreddit asyn
    posts = [] # List to store fetched post and comment data

    # Fetch top posts :
    async for post in subreddit.top(limit=num_posts):
        await post.load() # Load post data asynchronously
        comments = []

        # Load all comments
        await post.comments.replace_more(limit=0)

        # Define relevant keywords related to Rebel Sport products
        keywords = ['rebel sport', 'fitness', 'football', 'basketball', 'shoes',
                    'asics', 'new balance', 'under armour', 'nike', 'adidas',

        # Check if the comment or post body contains any of the keywords (return
        def contains_keywords(text):
            return any(keyword in text.lower() for keyword in keywords)

        # Collect the text of top 50 comments for the current post with keyword
        for comment in post.comments.list()[:50]:
            if isinstance(comment, asyncpraw.models.Comment) and contains_keyw
                comments.append(comment.body)

        # Store post data along with comments:
        posts.append({
            'subreddit': subreddit_name,
            'title': post.title or 'No Title',

```



```
        'score': post.score,  
        'id': post.id,  
        'url': post.url,  
        'num_comments': post.num_comments,  
        'created': post.created,  
        'body': post.selftext or 'No Content',  
        'comments': comments  
    })
```

```
    return posts
```

```
# Main async function to fetch and save posts from multiple subreddits
```

```
async def fetch_and_save_all():
```

```
    async with asyncpraw.Reddit(client_id=client_id,  
                                client_secret=client_secret,  
                                user_agent=user_agent) as reddit: # automatic session
```

```
    # Fetch posts from each subreddit:
```

```
    all_posts = [] # List to store all posts from different subreddits
```

```
    for subreddit in subreddits:
```

```
        posts = await fetch_posts_and_comments(reddit, subreddit)
```

```
        all_posts.extend(posts)
```

```
    # Convert the list of posts to a pandas DataFrame:
```

```
    df = pd.DataFrame(all_posts)
```

```
    # Save the DataFrame to a CSV file
```

```
    df.to_csv('/content/drive/MyDrive/Colab Notebooks/DWSMA/Assignment1/rebel_s
```

```
# Run the main async function using asyncio:
```

```
asyncio.run(fetch_and_save_all())
```

```
# Import library:
import pandas as pd

# Load the CSV file:
df = pd.read_csv('/content/drive/MyDrive/Colab Notebooks/DWSMA/Assignment1/rebe

# Display the first few rows of the DataFrame
print("First few rows of the DataFrame:")
print(df.head())

# Print the number of rows and columns of the DataFrame
print("\nNumber of rows and columns in the DataFrame:", df.shape)

# Print the column names of the DataFrame
print("\nColumn names of the DataFrame:", df.columns.tolist())
```



Show hidden output

## ✓ 5. C Data Wrangling

Clean text data, remove duplicates, handle non-standard characters, tokenize and normalise text, and classify sentiments.

This reduces unnecessary noise in the text data, making it more suitable for tokenization and sentiment analysis.

### 1. Text Cleaning (including standardization) and Preprocessing

The primary goal of text cleaning is to transform raw, unstructured text data into a format that is suitable for analysis. The steps include:

- **Handle missing value and duplicate value**
- **Removing special characters, hashtags, punctuation, and URLs:** These elements can clutter the data and don't contribute valuable meaning for most text analysis tasks.
- **Standardizing the text by:**
  - **Lowercasing:** Converts all text to lowercase to ensure consistency, eliminating case-sensitive discrepancies.
  - **Expanding contractions:** For example, changing "can't" to "cannot" ensures consistency and removes unnecessary variations.
  - **Removing extra spaces:** Ensures the text is formatted consistently, reducing the potential for errors in tokenization.
- **Removing stopwords:** Stopwords are common words like "the," "is," and "and" that are often filtered out in text analysis since they don't contribute much value to the meaning of the text.

```
# Duplicate entries can skew analysis results, leading to inaccurate conclusion
# Remove duplicate entries based on 'id'
df.drop_duplicates(subset=['id'], inplace=True)
# Check missing values before handling
# Handle missing values: Missing data can lead to biased or incomplete analysis
print("Missing values before handling:\n", df.isnull().sum())
```



Show hidden output

```
!pip install contractions
```



Show hidden output

```
import re
import nltk
from nltk.corpus import stopwords
from contractions import fix # A library to expand contractions

# Download stopwords from nltk
nltk.download('stopwords')
stop_words = set(stopwords.words('english'))

# Text cleaning and standardization function
def clean_and_standardize_text(text):
    # Handle NaN or non-string values
    if not isinstance(text, str):
        text = str(text)

    # Convert text to lowercase
    text = text.lower()

    # Expand contractions (e.g., "don't" -> "do not")
    text = fix(text)

    # Remove URLs
    text = re.sub(r"http\S+", "", text)

    # Remove special characters, numbers, and punctuation
    text = re.sub(r'[^a-zA-Z\s]', '', text)

    # Remove extra spaces
    text = re.sub(r'\s+', ' ', text).strip()

    # Remove stopwords
    text = " ".join([word for word in text.split() if word not in stop_words])

    return text

# Apply cleaning and standardization to 'body' column
df['body'] = df['body'].apply(clean_and_standardize_text)

# Handling 'comments' column
def clean_comments(comments):
    if isinstance(comments, str):
        try:
            # Evaluate the string into a list of comments
            comment_list = eval(comments)
            # Clean each comment in the list
            return [clean_and_standardize_text(comment) for comment in comment_
```

```
        except:
            return []
    return []

# Apply cleaning function to the 'comments' column
df['comments'] = df['comments'].apply(clean_comments)

# Check the first few rows after cleaning
print(df.head())
```



[Show hidden output](#)

## Explanation of Tasks

- **Removing Special Characters, Hashtags, Punctuation, and URLs:** Special characters, hashtags, punctuation, and URLs are often irrelevant to the analysis and can introduce unnecessary noise. For instance, a hashtag (e.g., #sports) or URL (e.g., [www.example.com](http://www.example.com)) does not convey significant sentiment information. Removing these elements helps focus on the actual content of the text.
- **Standardization:** Lowercasing ensures uniformity in the text, preventing discrepancies such as "Sports" and "sports" being treated as different terms. Expanding contractions like "don't" to "do not" standardizes words that may have multiple forms, ensuring consistency across the dataset. Removing extra spaces ensures there are no inconsistencies in text formatting, which can disrupt further processing tasks.
- **Removing Stopwords:** Stopwords are words that occur frequently in a language but add little to the meaning of a sentence. They can distract the algorithms from identifying the key themes or sentiments in the text. By eliminating these words, the focus is placed on more meaningful, content-driven terms, which is especially important for tasks like sentiment analysis or topic modeling.

## Justification of Wrangling Decisions

- **Focus on Meaningful Content:** Removing special characters, punctuation, URLs, and stopwords helps the algorithm focus on the meaningful content of the text. Without these distractions, the model can better identify the important terms and their relationships, leading to more accurate and insightful results.
- **Improved Analysis Efficiency:** By standardizing text, cleaning unwanted characters, and removing redundant terms like stopwords, the dataset becomes smaller and more manageable. This allows for faster processing and analysis, improving computational efficiency, particularly for large datasets.
- **Enhanced Accuracy in Analysis:** Cleaning the text and removing noise ensures that the sentiment analysis or any other text classification task is based on the most relevant words. This leads to more accurate sentiment classification because the model isn't distracted by irrelevant information.
- **Better Focus on Core Sentiment:** In sentiment analysis, words such as "good," "happy," "love," and "hate" carry more weight in determining the sentiment of a text. Removing stopwords or irrelevant characters ensures that these key terms receive the attention they deserve, which improves sentiment classification accuracy.

## ✓ 2. Tokenization and Normalization

Tokenization and normalization are crucial tasks that convert the cleaned text into a structured format that can be analyzed. Tokenization involves breaking the text into individual units, or tokens, typically words or phrases, while normalization involves standardizing these tokens to make them consistent.

- **Tokenization:** Split the cleaned text into individual words or phrases (tokens) for further analysis.
- **Normalization:** Includes processes like stemming (reducing words to their root form) or lemmatization (converting words to their base form) to ensure consistency in the representation of terms.

```
from nltk.stem import WordNetLemmatizer
nltk.download('wordnet')

lemmatizer = WordNetLemmatizer()

# Lemmatize the tokenized words in comments and body
df['body'] = df['body'].apply(lambda text: " ".join([lemmatizer.lemmatize(word)
df['comments'] = df['comments'].apply(lambda comments: [" ".join([lemmatizer.le
                                for word in comment.split())) for comment in comments])
```



Show hidden output



## Explanation of Tasks

- **Tokenization:** This breaks down the text into individual components (tokens), which can then be analyzed for frequency, sentiment, and relationships. Tokenization is essential because text is unstructured, and breaking it down into individual pieces is the first step in making it analyzable.
- **Normalization:** This step ensures that different forms of the same word are treated as identical. For example, the words "running" and "runner" could be normalized to the base form "run" to avoid redundancy and inconsistency in the analysis. Lemmatization typically involves more complex language rules compared to stemming, ensuring that only meaningful base forms are considered.

## Justification of Wrangling Decisions

- **Standardization for Accurate Representation:** Tokenization is essential for breaking down complex text into manageable chunks. By normalizing the tokens, the analysis treats different variations of words as the same, ensuring that the model recognizes them as related concepts. For example, "running" and "run" would be treated as equivalent, allowing for a more accurate understanding of text patterns.
- **Facilitating Analysis:** By converting the text into tokens and normalizing them, it becomes easier to apply algorithms like sentiment analysis, frequency analysis, and topic modeling. These tasks require a consistent and structured representation of text to identify patterns or relationships between terms.

## ✓ 5. D Findings from Reddit Data Analysis

The analysis of Reddit discussions provides valuable insights into customer preferences, popular products, and emerging market trends. The visualizations below highlight the most commonly mentioned products, brands, and keywords, shedding light on customer interests and engagement.

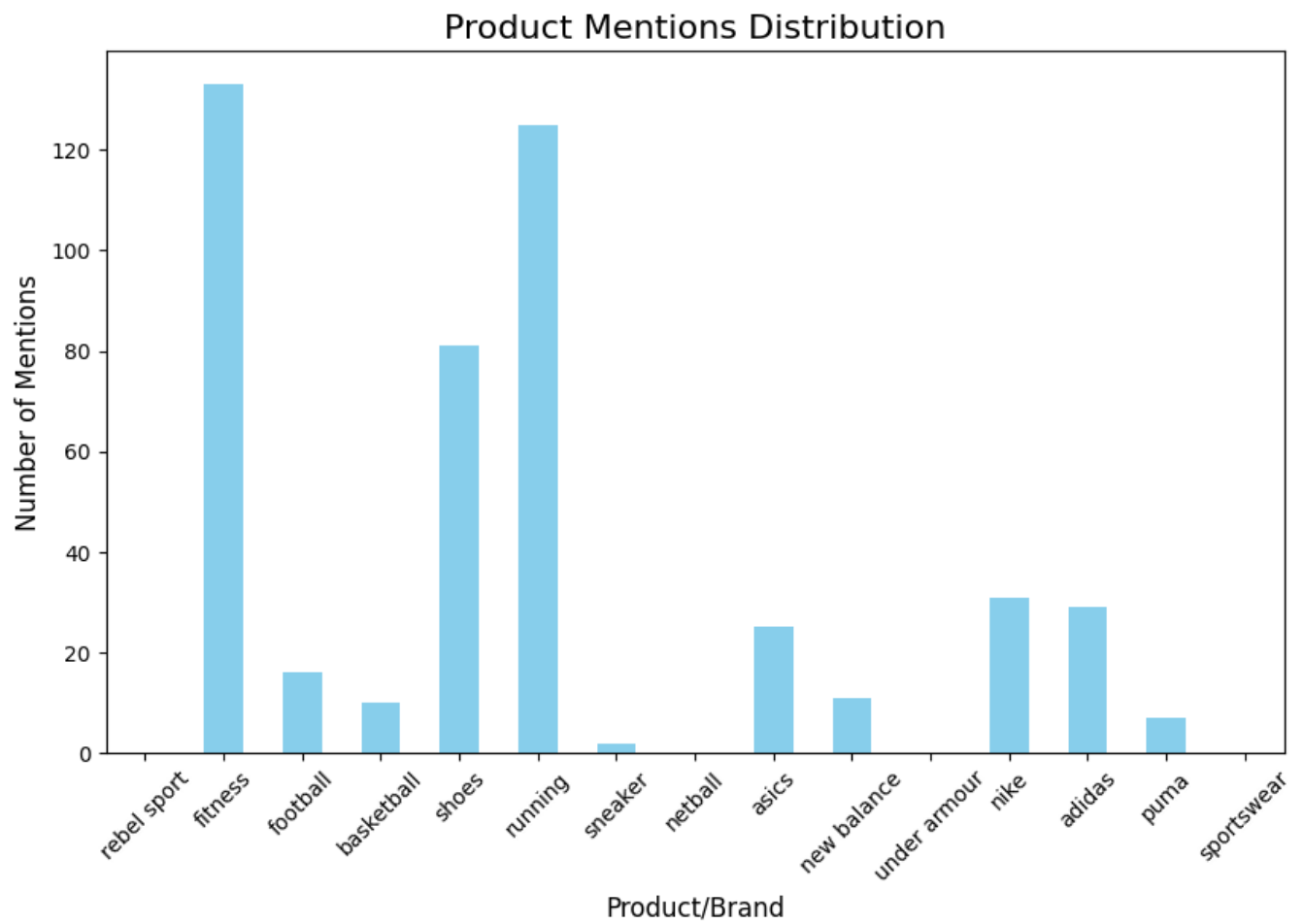
### 1. Product and Brand Mentions (Bar Chart)

This chart demonstrates the frequency of product and brand mentions on Reddit.

#### Key Insights:

- Products related to **fitness, running gear, and shoes** dominate the discussions, reflecting high user engagement in these categories.
- **Top mentioned brands** include **Nike, Adidas, and Asics**, indicating strong consumer interest and loyalty toward these brands.
- **Implication:** Rebel Sport can optimize its inventory and marketing efforts by focusing on these high-demand categories and popular brands.

```
from IPython.display import Image, display
img_path = '/content/drive/MyDrive/Colab Notebooks/DWSMA/Assignment1/1a.png'
display(Image(filename=img_path))
```



- 2. Word Cloud: Frequently Mentioned Keywords

This word cloud highlights the most frequently mentioned terms in Reddit discussions.

### Key Insights:

- Prominent words include “**shoe,**” “**running,**” “**fitness,**” “**Nike,**” and “**gym,**” indicating a strong association with footwear and fitness-related products.
- Other frequently mentioned terms, such as “**pretty,**” “**size,**” and “**feel,**” suggest that users are actively discussing product aesthetics, fit, and comfort.
- **Implication:** These findings present an opportunity for Rebel Sport to refine its product recommendations and emphasize key attributes valued by customers, such as comfort, style, and functionality.

```
img_path = '/content/drive/MyDrive/Colab Notebooks/DWSMA/Assignment1/Unknown-7.  
display(Image(filename=img_path))
```



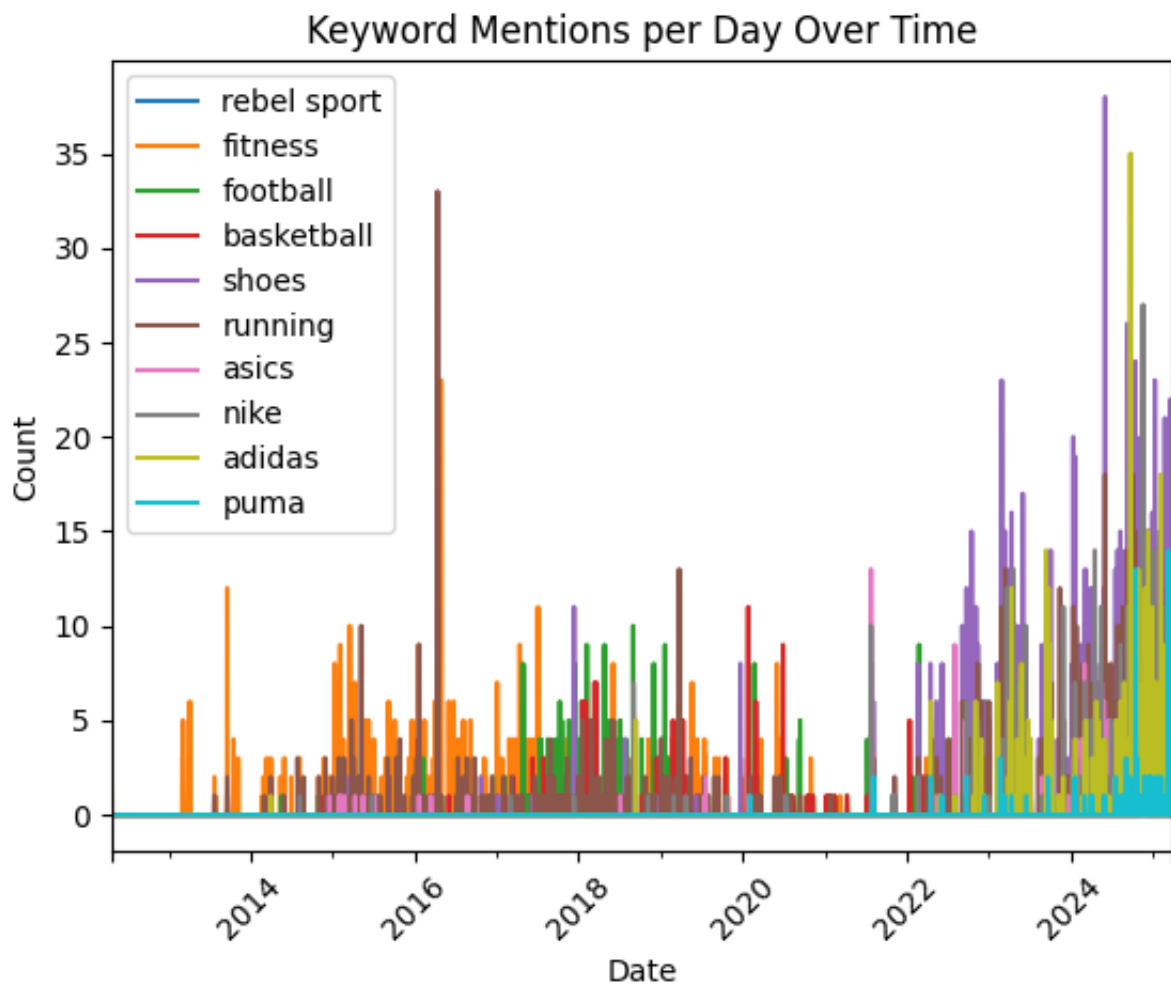
### ✓ 3. Time Series Plot: Trending Mentions Over Time

This plot tracks the daily mentions of key products and brands over time.

#### Key Insights:

- A **recent upward trend** is observed in mentions of **shoes, running, Nike, Adidas, and Puma**, indicating growing consumer interest in these products and brands.
- Possible drivers: Seasonal factors, product launches, or events may be contributing to the surge in mentions.
- **Implication:** By tracking these trends in real-time, Rebel Sport can align product restocking and promotional campaigns with periods of increased consumer interest.

```
img_path = '/content/drive/MyDrive/Colab Notebooks/DWSMA/Assignment1/Unknown-6.  
display(Image(filename=img_path))
```



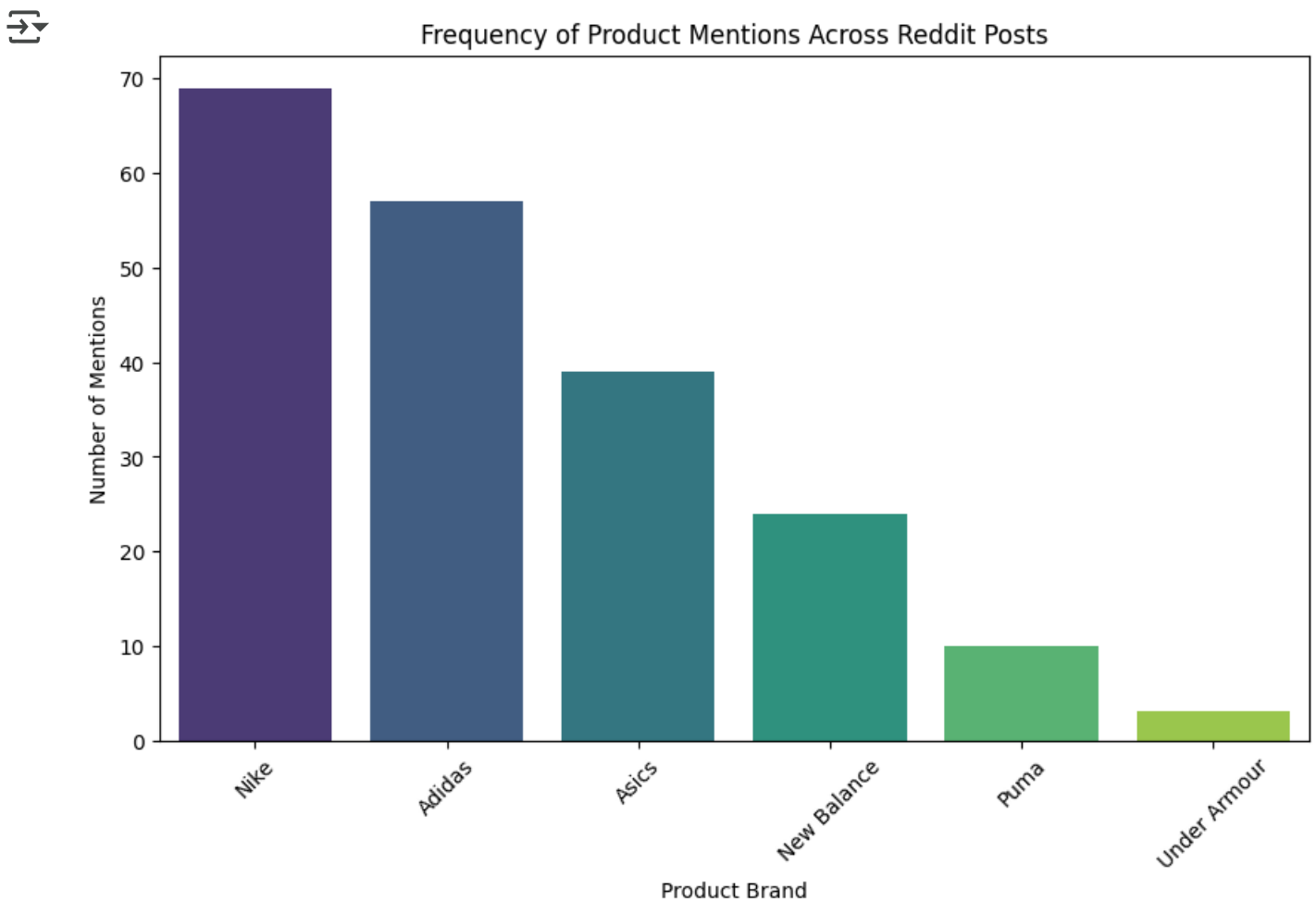
#### ✓ 4. Top Three Product Mentions (Bar Chart)

This bar chart presents the most frequently mentioned products and brands across Reddit.

##### Key Insights:

- Nike is the most frequently mentioned brand, followed by Adidas and Asics.
- The high frequency of mentions reflects strong brand loyalty and customer preference.
- **Implication:** Prioritizing these brands in marketing campaigns and product promotions can enhance customer satisfaction and drive higher sales.

```
img_path = '/content/drive/MyDrive/Colab Notebooks/DWSMA/Assignment1/Unknown-8.  
display(Image(filename=img_path))
```



Connection to Rebel Sport's Proposal: Personalization through Rebel

## GearMatch

The insights from Reddit align closely with the goals of the Rebel GearMatch tool, which aims to enhance product recommendations through personalization.

- **Personalized Recommendations:** The analysis reveals that customers value product-specific attributes such as comfort, size, and style. Incorporating these preferences into the Rebel GearMatch tool can improve the relevance and accuracy of product comparisons, enhancing the user experience.
- **Real-Time Trend Tracking:** The observed trends in product mentions can help Rebel Sport stay ahead of shifting customer preferences, ensuring that the Rebel GearMatch tool remains up-to-date with the latest market dynamics.
- **Brand Loyalty and Engagement:** The strong engagement with brands like Nike and Adidas indicates that Rebel GearMatch should highlight popular brands and trending products to boost customer engagement and satisfaction.

## Persuasiveness of the Argument: Supporting the Application's Success

The positive feedback and high engagement levels observed on Reddit suggest strong market potential for the Rebel GearMatch tool.

- **Data-Driven Personalization Boosts Customer Loyalty:** Research shows that personalized shopping experiences can increase customer satisfaction and retention. By leveraging Reddit data, Rebel Sport can deliver tailored recommendations that meet users' actual needs.
- **Enhanced Shopping Experience:** The insights on trending products and user preferences support Rebel Sport's strategy to provide a seamless, customized, and data-driven shopping journey.
- **Competitive Advantage:** By aligning product offerings with real-time trends and customer sentiment, Rebel Sport can differentiate itself in a competitive market, ultimately driving sales growth and long-term profitability.

## Conclusion

The Reddit data analysis provides actionable insights that reinforce the potential success of the Rebel GearMatch tool. By leveraging these findings, Rebel Sport can enhance personalization, improve customer engagement, and stay ahead of market trends, leading to sustainable business growth.



# References

1. Annual report, (2024). Supper Retail Group,  
<https://media.supercheapauto.com.au/corp/files/documents/2024%20Annual%20Report.pdf>
2. Backlinko Team, (2025). Backlinko. <https://backlinko.com/reddit-users>
3. Koen van Gelder, (2024). Statista. <https://www.statista.com/statistics/1239795/price-comparison-online-shopping-habits/>
4. Jose, H. (2023). Forbes.  
<https://www.forbes.com/councils/forbesbusinesscouncil/2023/06/28/stay-connected-with-customers-through-ultra-personalized-experiences/>
5. McKinsey & Company. (2021). "The Value of Personalized Marketing." McKinsey & Company. Available: <https://www.mckinsey.com>
6. Bain & Company, 2024. The value Of Online customer loyalty and how you can capture it. [https://media.bain.com/Images/Value\\_online\\_customer\\_loyalty\\_you\\_capture.pdf](https://media.bain.com/Images/Value_online_customer_loyalty_you_capture.pdf)