Big Data Analytics in Engineering

Final Project Showcase

Topic: Enhance Business Strategy with Sentiment Analysis

Sub-topic: Data-Driven Insights for Promotion Optimization

Dataset: Walmart-extended-dataset

By: Partha Sarathi Kundu

WSUID: XXXXXXX

Introduction

Objective:

- Explore the role of sentiment analysis and satisfaction scores in predicting customer purchase likelihood.
- Guide business strategy on promotions using data-driven insights.

Scope:

- Analyze customer reviews, satisfaction scores, and purchase patterns.
- Develop actionable strategies based on sentiment trends.
- Design Future-proof dynamic data-driven solution.

Data Overview

- **Dataset Description:**
  - 3 different datasets, 33 combined columns
  - 55,000 records of customer reviews, purchases, promotion and satisfaction scores.
  - Key variables: `Sentiments (Review & Reply)`, `DiscountPercentage`, `EngagementMetric`, `CustomerInteraction`, `ConditionalPromotions` and `Purchase`.
- **Goal:**
  - Perform extensive EDA, Feature enrichment and data Normalization
  - Analyse customer sentiments to identify trends
  - Purchase behavior based on Sentiment analysis
  - Correlations to optimize promotion strategies for business profit

Data Overview [Con.]

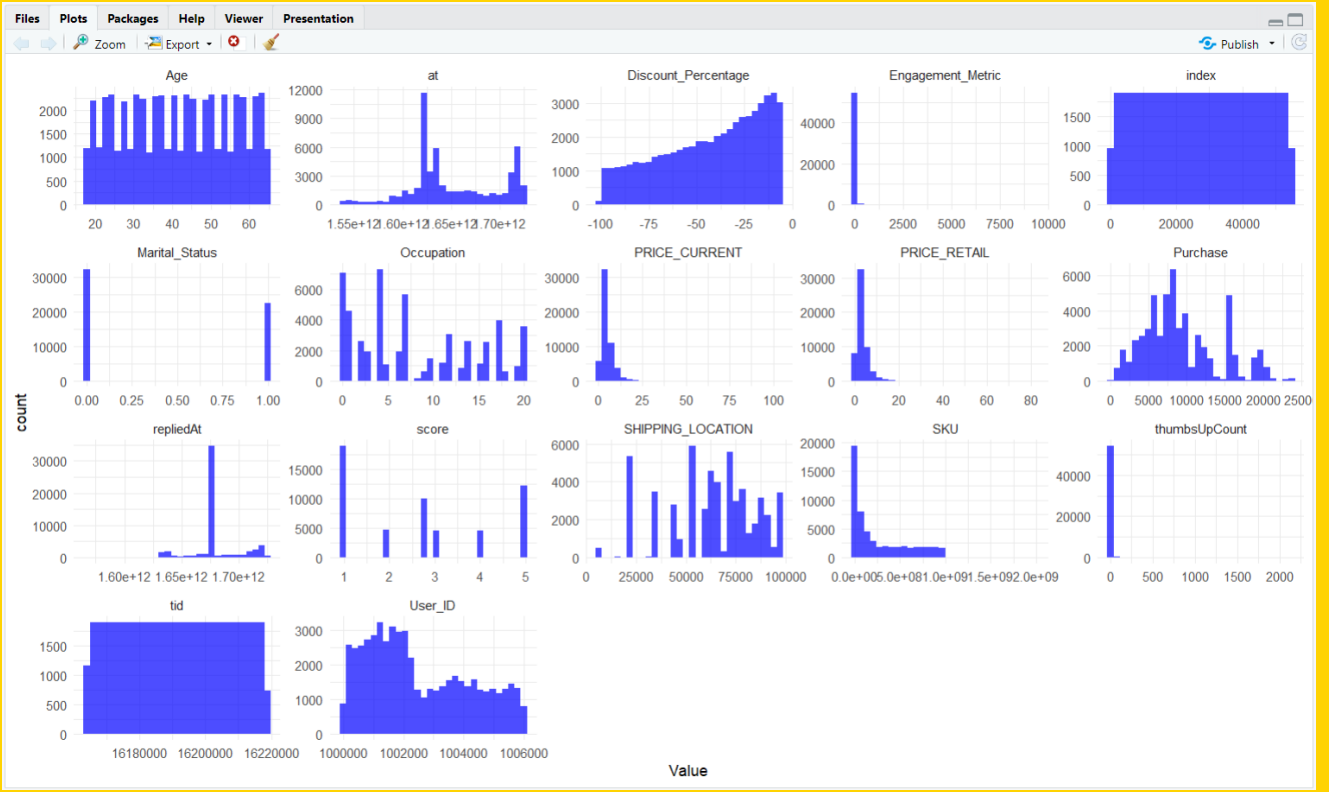
```
Console  Terminal  Background Jobs
R 4.4.2 · C:/Users/t3student7/Downloads/walmart/refineddataset/mergeddataset/
> # Load the dataset (replace with the actual path to your file)
> final_dataset <- read.csv("finaldataset.csv") # Ensure "finalDataset.csv" is in your working directory
> # 1. View the structure and summary of the dataset
> str(final_dataset)
'data.frame':   55000 obs. of  33 variables:
 $ index                  : int  0 1 2 3 4 5 6 7 8 9 ...
 $ SHIPPING_LOCATION      : int  79936 79936 79936 79936 79936 79936 79936 79936 79936 79936 ...
 $ DEPARTMENT             : chr  "Deli" "Deli" "Deli" "Deli" ...
 $ CATEGORY               : chr  "Hummus, Dips, & Salsa" "Hummus, Dips, & Salsa" "Hummus, Dips, & Salsa"
 $ SUBCATEGORY            : chr  "Dips & Sauces" "Dips & Sauces" "Dips & Sauces" "Dips & Sauces" ...
 $ BREADCRUMBS            : chr  "Deli/Hummus, Dips, & Salsa" "Deli/Hummus, Dips, & Salsa" "Deli/Hummus,
 $ SKU                    : int  110895339 105455228 128642379 366126367 160090316 174071300 10294995 37
d-Garlic-Hummus-10-Oz/105455228?fulfillmentIntent=Pickup" "https://www.walmart.com/ip/Marketside-Classic-Hu
side-Everything-Hummus-10-oz/366126367?fulfillmentIntent=Pickup" ...
 $ PRODUCT_NAME           : chr  "Marketside Roasted Red Pepper Hummus, 10 Oz" "Marketside Roasted Garli
s, 10 oz" ...
 $ BRAND                  : chr  "Marketside" "Marketside" "Marketside" "Marketside" ...
 $ PRICE_RETAIL           : num  1.38 2.42 2.25 1.53 2.75 ...
 $ PRICE_CURRENT          : num  2.67 2.67 2.67 2.67 3.12 3.12 2.42 5.54 3.27 4.54 ...
 $ PRODUCT_SIZE           : chr  "10" "10" "10" "10" ...
 $ PROMOTION              : chr  "Yes" "No" "Yes" "Yes" ...
 $ tid                    : int  16163804 16163805 16163806 16163807 16163808 16163809 16163810 16163811
 $ User_ID                : int  1000001 1000001 1000001 1000001 1000002 1000003 1000004 1000004 1000004
 $ Gender                 : chr  "F" "F" "F" "F" ...
 $ Age                    : int  50 34 64 33 27 46 59 48 34 39 ...
 $ Occupation             : int  10 10 10 10 16 15 7 7 7 20 ...
 $ City_Category          : chr  "A" "A" "A" "A" ...
 $ Stay_In_Current_City_Years: chr  "2" "2" "2" "2" ...
 $ Marital_Status         : int  0 0 0 0 0 0 1 1 1 1 ...
 $ Purchase               : int  8370 15200 1422 1057 7969 15227 19215 15854 15686 7871 ...
 $ CustomerReviewContent  : chr  "Fantastic app! It's so easy to use, and I rarely have issues with the
y going to be helpful. I occasionally need a physical receipt. Then, the other day, I got"| __truncated__ "
lmart has, they can't hire som"| __truncated__ "apps design is pretty good. But I can't give 5 stars becaus
 $ score                  : num  4 3 3 3 3 2 1 1 1 3 ...
 $ thumbsUpCount          : num  153 18 68 11 9 28 13 15 21 11 ...
 $ at                     : num  1.72e+12 1.72e+12 1.72e+12 1.72e+12 1.72e+12 1.72e+12 ...
 $ WalmatReplyContent     : chr  "We value our customers review. Please provide us with more info at htt
your receipts effectively, and we value your input. Please s"| __truncated__ "We understand the frustration
"| __truncated__ "We're glad you like the app's design but understand your frustration with the technical i
 $ repliedAt              : num  1.72e+12 1.72e+12 1.72e+12 1.72e+12 1.72e+12 ...
 $ appName                : chr  "Walmart" "Walmart" "Walmart" "Walmart" ...
 $ Product_ID             : chr  "P00069042" "P00248942" "P00087842" "P00085442" ...
 $ Discount_Percentage    : num  -93.6 -10.2 -18.9 -74.4 -13.4 ...
 $ Engagement_Metric      : num  612 54 204 33 27 56 13 15 21 33 ...
```

```
Console  Terminal  Background Jobs
R 4.4.2 · C:/Users/t3student7/Downloads/walmart/refineddataset/mergeddataset/
> # Correlation matrix for numeric variables
> correlation_matrix <- cor(final_dataset %>% select_if(is.numeric), use = "complete.obs")
> print(correlation_matrix)
                          index SHIPPING_LOCATION          SKU PRICE_RETAIL PRICE_CURRENT          tid
index              1.0000000000      0.1178592581  0.0197528742 -0.005410558 -0.0064059816  1.0000000000
SHIPPING_LOCATION  0.1178592581      1.0000000000  0.0456367361  0.033493449  0.0362257562  0.1178592581
SKU                0.0197528742      0.0456367361  1.0000000000  0.092536094  0.0956435950  0.0197528742
PRICE_RETAIL      -0.0054105579      0.0334934490  0.0925360943  1.000000000  0.9600176717 -0.0054105579
PRICE_CURRENT     -0.0064059816      0.0362257562  0.0956435950  0.960017672  1.0000000000 -0.0064059816
tid                1.0000000000      0.1178592581  0.0197528742 -0.005410558 -0.0064059816  1.0000000000
User_ID            0.1738130052     -0.1255607748  0.0143283121 -0.002779980 -0.0042648785  0.1738130052
Age               -0.0001253607      0.0047100296 -0.0063641603  0.002184657  0.0013824805 -0.0001253607
Occupation        -0.0218492045     -0.0153439579  0.0064989716  0.004962902  0.0065442238 -0.0218492045
Marital_Status     0.0195398617     -0.0014485738  0.0038547594  0.003910082  0.0056646223  0.0195398617
Purchase           0.0030108488      0.0133793540  0.0096043137 -0.001624679 -0.0002834393  0.0030108488
score              0.0131528822      0.0163616652 -0.0081364918  0.003311450  0.0049284885  0.0131528822
thumbsUpCount     -0.0670582425      0.0149444793  0.0053837117  0.006077425  0.0059884574 -0.0670582425
at                -0.4310623154     -0.0048687247 -0.0179642761  0.012976607  0.0147741073 -0.4310623154
repliedAt         -0.2748406492      0.0298835800 -0.0160907510  0.016296625  0.0178812616 -0.2748406492
Discount_Percentage  0.0011525292   -0.0004529438  0.0007667103  0.218860839 -0.0015182316  0.0011525292
Engagement_Metric -0.0389913878      0.0151292348  0.0034115164  0.005605309  0.0059011853 -0.0389913878
                        User_ID           Age   Occupation Marital_Status      Purchase         score
index              0.173813005 -0.0001253607 -0.021849204   1.953986e-02  3.010849e-03  0.013152882
SHIPPING_LOCATION -0.125560775  0.0047100296 -0.015343958  -1.448574e-03  1.337935e-02  0.016361665
SKU                0.014328312 -0.0063641603  0.006498972   3.854759e-03  9.604314e-03 -0.008136492
PRICE_RETAIL      -0.002779980  0.0021846571  0.004962902   3.910082e-03 -1.624679e-03  0.003311450
PRICE_CURRENT     -0.004264878  0.0013824805  0.006544224   5.664622e-03 -2.834393e-04  0.004928489
tid                0.173813005 -0.0001253607 -0.021849204   1.953986e-02  3.010849e-03  0.013152882
User_ID            1.000000000 -0.0050185666 -0.026919632   1.367790e-02  5.398804e-03  0.025742739
Age               -0.005018567  1.0000000000  0.003223102   5.675691e-04 -8.323034e-03 -0.002848071
Occupation        -0.026919632  0.0032231018  1.000000000   1.078525e-02  1.261865e-02  0.004183680
Marital_Status     0.013677903  0.0005675691  0.010785246   1.000000e+00 -9.529897e-05 -0.001776838
Purchase           0.005398804 -0.0083230341  0.012618646  -9.529897e-05  1.000000e+00 -0.001945491
score              0.025742739 -0.0028480715  0.004183680  -1.776838e-03 -1.945491e-03  1.000000000
thumbsUpCount     -0.059933406  0.0038321892  0.001128584  -9.792235e-03  3.288976e-03 -0.032552923
at                -0.074514570  0.0025959804  0.012676859  -1.608896e-02  8.687086e-04  0.314213365
repliedAt         -0.138594847 -0.0007263653  0.006782218  -1.412548e-02 -6.286055e-03  0.284494834
Discount_Percentage  0.007024194  0.0072734880 -0.009319419  -1.659544e-03 -7.747406e-03 -0.006874374
Engagement_Metric -0.039515667  0.0011152934 -0.001698975  -9.211626e-03  5.052521e-03  0.034426584
                  thumbsUpCount            at     repliedAt Discount_Percentage Engagement_Metric
index             -0.067058242 -0.4310623154 -0.2748406492        0.0011525292       -0.038991388
SHIPPING_LOCATION  0.014944479 -0.0048687247  0.0298835800       -0.0004529438        0.015129235
SKU                0.005383712 -0.0179642761 -0.0160907510        0.0007667103        0.003411516
PRICE_RETAIL       0.006077425  0.0129766065  0.0162966251        0.2188608392        0.005605309
PRICE_CURRENT      0.005988457  0.0147741073  0.0178812616       -0.0015182316        0.005901185
```

WICHITA STATE UNIVERSITY

Data Overview [Con.]

```
> # Check for missing values
> missing_values <- colSums(is.na(final_dataset))
> print(missing_values)
                    index         SHIPPING_LOCATION                DEPARTMENT                  CATEGORY
                        0                         0                         0                         0
              SUBCATEGORY               BREADCRUMBS                       SKU               PRODUCT_URL
                        0                         0                         0                         0
             PRODUCT_NAME                     BRAND              PRICE_RETAIL             PRICE_CURRENT
                        0                         0                         0                         0
             PRODUCT_SIZE                 PROMOTION                       tid                   User_ID
                        0                         0                         0                         0
                   Gender                       Age                Occupation             City_Category
                        0                         0                         0                         0
Stay_In_Current_City_Years            Marital_Status                  Purchase     CustomerReviewContent
                        0                         0                         0                         0
                    score             thumbsUpCount                        at          WalmatReplyContent
                        0                         0                         0                         0
                repliedAt                   appName                Product_ID         Discount_Percentage
                        0                         0                         0                         0
       Engagement_Metric
                        0
```



Key Points:

Uniform Distributions:

- Variables like **Age** and **Marital_Status** show uniform or binary distributions, indicating diverse demographics and clear categories.

Skewed Variables:

- Variables like **PRICE_CURRENT**, **thumbsUpCount**, and **Engagement_Metric** are highly skewed, with most values near zero, indicating limited high-value interactions or outliers.

Purchase Patterns:

- **Purchase** shows a bell-shaped distribution with a peak in the mid-range, suggesting most purchases cluster around average amounts.

Discount Impact:

- Discounts are heavily skewed toward significant negative values, highlighting aggressive pricing strategies.

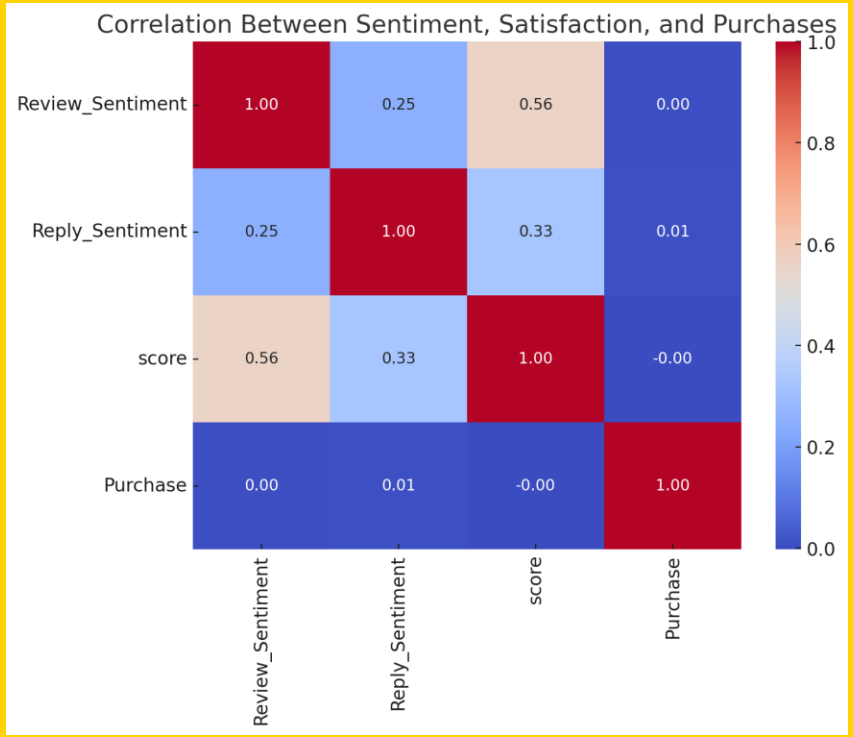
Satisfaction (score):

- Scores are discrete and lean toward mid-to-high values, indicating general customer satisfaction.

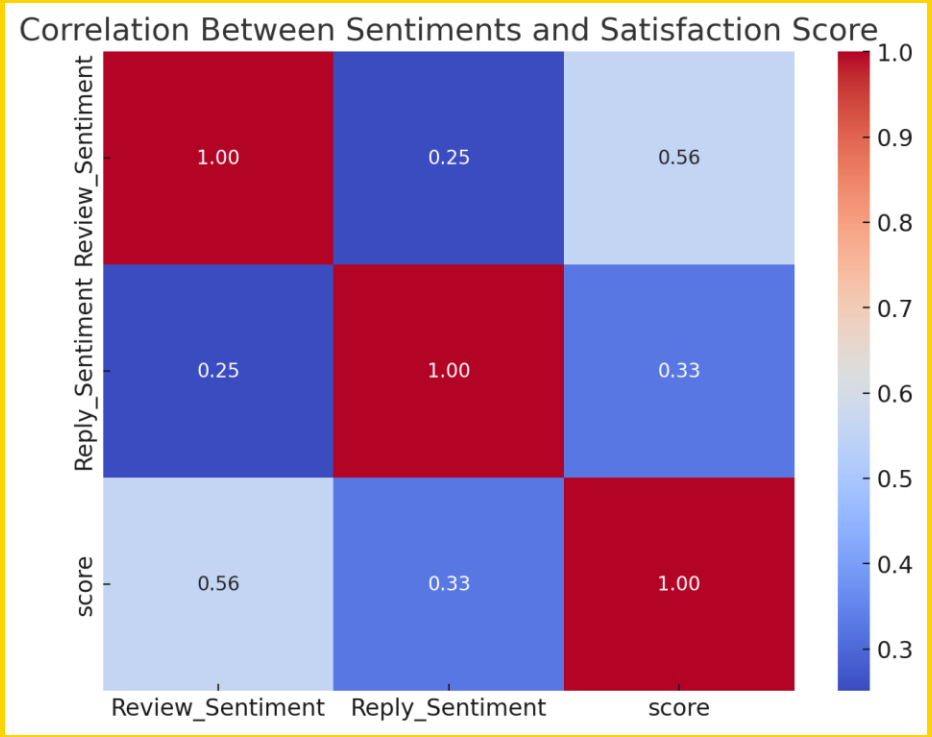
WICHITA STATE UNIVERSITY

Key Analysis

Correlation Heatmap between features customer sentiments and interaction score



Takeaway: Satisfaction and sentiment indirectly influence purchases but play a role in engagement.



Takeaway: Positive reviews are moderately linked to higher customer satisfaction (correlation: 0.56). Replies from Walmart have a weaker impact on satisfaction (correlation: 0.33) compared to reviews.

WICHITA STATE UNIVERSITY

Key Analysis [Con.]

Review text mining and WordCloud, Text Mining for Sentiment analysis

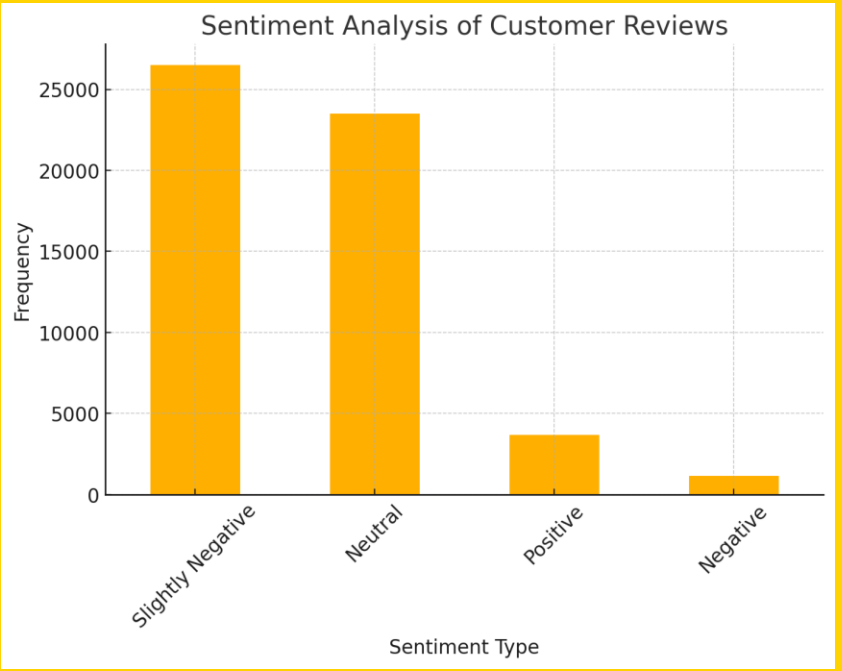


Fig: Customer review text mining and Positive/Neutral Sentiment Word Cloud

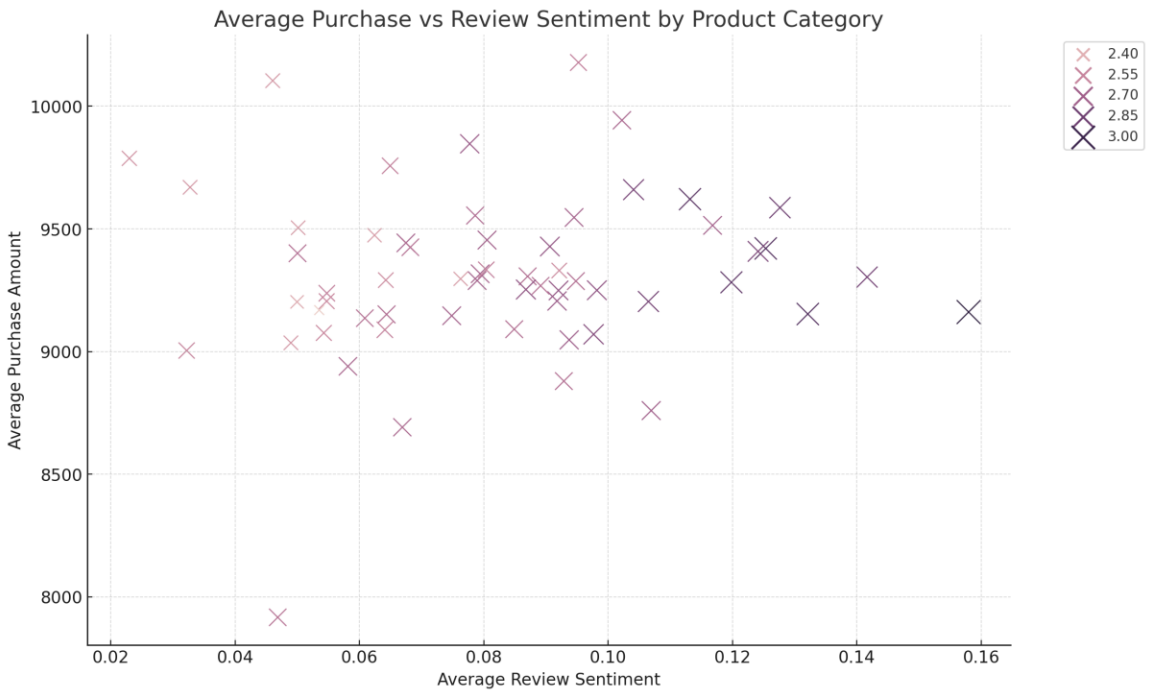
Observations:

- The majority of reviews lean towards neutral or slightly negative tones, suggesting mixed customer experiences.
- Positive reviews are less common, indicating potential gaps in customer satisfaction or areas where improvements could boost favorable feedback.

WICHITA STATE UNIVERSITY

Key Analysis [Con.]

Relationship between Sentiment and Purchase Trends

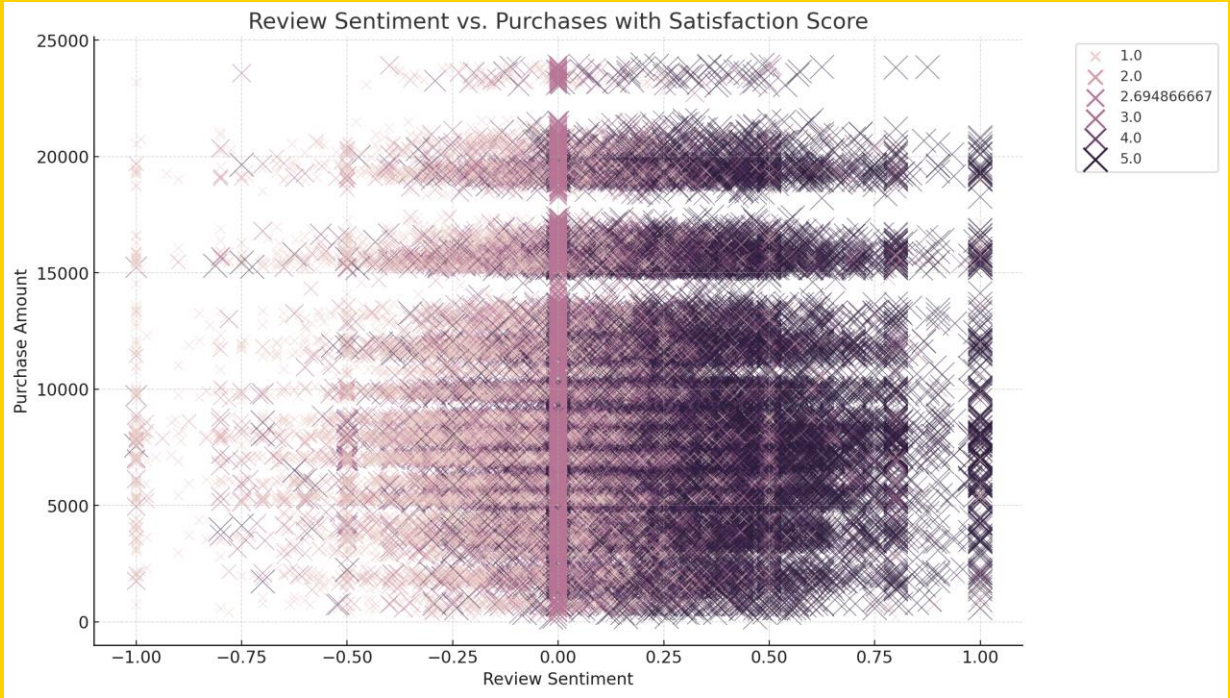


Observation:

- As review sentiment becomes more positive, average purchase amounts tend to increase for some product categories.
- Categories with higher review sentiment are associated with slightly higher satisfaction scores.

Key Takeaway:

- Positive customer sentiment aligns with higher purchases in specific categories, indicating sentiment as a driver for purchase behavior.



Observation:

- The plot shows a wide distribution of purchase amounts across the range of review sentiments.
- Neutral sentiment (close to 0) has significant overlap with varying purchase levels, while extremely positive or negative sentiments show fewer data points. Satisfaction scores (size of points) increase slightly with more positive sentiments.

Key Takeaway:

- While positive sentiments may slightly align with higher purchase amounts, the impact is less distinct for neutral or slightly negative sentiments.

WICHITA STATE UNIVERSITY

Data driven Strategy and Decision

Average Purchases by Promotion Strategy

Observation:
Products categorized into:
- **Promote:** High sentiment, high satisfaction, and above-average purchases.
- **No Promotion:** Lower sentiment or satisfaction.

Results: "Promote" products show significantly higher purchase amounts.
Action Point: Allocate resources to promote high-sentiment products to maximize ROI.

Target Products for Promotion:
- Focus on high-sentiment, high-satisfaction products with untapped potential.

Address Neutral/Negative Sentiment:
- Use promotions to re-engage customers and improve sentiment.

Optimize Messaging:
- Tailor promotional campaigns based on customer feedback (e.g., emphasize affordability or quality).

Track Promotion Effectiveness:
- Use sentiment analysis pre- and post-promotion to measure success.

Business Strategy Guidance:
- **Promotion Focus:** Target products in the "Promote" category. Promotions can enhance existing customer engagement and maximize ROI.
- **Improve Strategy for "No Promotion" Products:** Investigate why these products underperform (e.g., lower satisfaction or sentiment) and address root causes before applying promotions.

WICHITA STATE UNIVERSITY

Business Impact & Future Improvements

Streamline benefits of Sentiment-Based Promotions and Automation

Business Benefits:

- Improved customer engagement.
- Increased sales for underperforming yet positively reviewed products.
- Enhanced customer trust and loyalty through personalized strategies.

Boost Revenue with Sentiment-Based Promotions:

- Target high-sentiment products for maximum ROI and address low-sentiment products to re-engage customers.

Enhance Customer Trust and Retention:

- Use sentiment insights to resolve customer concerns and deliver personalized, impactful promotions.

Automation is the Future!

Automate Sentiment Analysis for Efficiency:

- Real-time monitoring and predictive analytics streamline promotions and resource allocation.
- Create Dynamic Promotion Strategies.
- Integrated Feedback Loops for 2-way communication.
- Cross-Channel Personalization for accessibility and inclusion.
- Deploy Agentic AI to reduce waste on ineffective campaigns and promotions.

WICHITA STATE UNIVERSITY

Data References & Code Repository

- **Data Reference**
  - https://www.kaggle.com/datasets/thedevastator/product-prices-and-sizes-from-walmart-grocery
  - https://www.kaggle.com/datasets/devarajv88/walmart-sales-dataset
  - https://www.kaggle.com/datasets/bleuarmendariz/walmart-employee-reviews
  - finaldataset.csv
- **Code Repo**
  - finalCode

Thank you all.

Feedback, Questions & Discussion.