Project Deliverable 3 - Team A6

1. Exploratory Data Analysis

Transaction Data: We extracted transaction data from Six Squad SQL server. We found that some columns needed to be preprocessed (string DataAndTime to datetime, ProductName from Russian to English). We discovered several outliers by price, discount and quantity, and verified that all these data points represented anomalous, but valid transactions. We found that most purchases involved a single item, with the average price of products roughly $140 USD. Sales are growing faster at the newer store, with weekend upticks in net sales across both locations.

NLP Data: Exploring this dataset, we can observe that it contains 13,595 rows, with each row containing 5 attributes: a review, the personal profile of the reviewer, the number of likes this review has received, the store the review is on, and the region of the store. Because the 'Personal profile' column contains multiple pieces of information, it is necessary to separate them during the pre-processing phase. Moreover, both the 'Personal profile' and 'Review' columns contain various irrelevant characters, which also requires attention during the pre-processing steps.

2. Pre - processing

Transaction data: We fixed the previously identified data type and language issues, and handled pre-processing separately for each technique employed -

- **Market Basket Analysis** - The transaction data was converted into a list format, with each list containing items purchased in a transaction. Then we transformed the data to the appropriate format using TransactionEncoder - 4,545 transactions, 1,082 distinct items.
- **For Clustering -** The transaction data was grouped and aggregated by CustomerID to get customer-level purchase history, and the numerical attributes were standardized.

NLP Data: We first used Regex to remove all non-alpha-numeric characters. Next, we transformed the 'Personal profile' into three new attributes. One is a binary column indicating whether the reviewer is a local guide or not. The other two are numerical variables showing the total number of reviews the reviewer has left and the total number of photos the reviewer has uploaded. Finally, we normalized words and reduced them to their root forms for the 'Review'.

3. Analysis Plan

Transaction data: We opted for two types of unsupervised ML methods -

- **MBA** - We wanted to understand the purchase behavior of the customers, so the insights can be used for various applications such as cross-selling, store layout and targeted marketing. We have used mlxtend.frequent_patterns library to uncover associations between items and identify purchasing patterns within Six Squad's transactions.
- **Clustering** - Our goal is to identify distinct customer segments in Six Squad's loyalty program. We will test 4 clustering algorithms - KMeans (general purpose), Hierarchical (effective at identifying non-globular clusters), KPrototypes (can handle categorical data), and DBSCAN (effective at identifying clusters by density). We will also test out preprocessing & tuning methods (best K, PCA to reduce noise, optimal parameters etc.). Silhouette score will be our main metric for evaluation.

NLP Data: We scraped reviews from 45 sneaker stores in pop culture trendsetting cities to pinpoint key aspects making them trendy. Using TextBlob for an initial polarity overview, we ensured an ample number of positive or neutral reviews. Employing Latent Dirichlet Allocation

(LDA) for topic modeling, we identified topics and associated words with weights. BERT-base-uncased model processed LDA's crucial words and our researched terms, yielding sentiment scores for each comment. We calculated average sentiment scores at the store level and listed the top 3 stores per aspect, providing valuable recommendations for Six Squad shadowing.

4. Results

Transaction data: We have the following preliminary results -

- **MBA** -"WMNS AIR JORDAN 1 LOW" is the most popular product, appearing in 2.27% of transactions. Association rule analysis showed a strong trend of purchasing series items together, like a 35.71% chance of buying "Bearbrick Blindbox series 46" after series 45. Time-based analysis revealed significant shopping behavior differences, with 558 items frequently bought on weekends versus only 10 on weekdays, indicating distinct consumer preference for shopping on weekends. From our analysis at category level, we observed that customers tend to purchase outfits rather than individual clothing items. But when they aim to buy individual items, sneakers are most often chosen. Based on these insights, we have made three slight changes to our store layout (please refer to the picture in appendix), to positively influence customers' purchase decisions:
    - Move the changing room next to the sneakers display. That way, customers who intend to purchase outfits will be exposed to sneakers (outfits + sneakers have high lift).
    - Relocate caps section to the cashier counter to also increase their visibility.
    - Add two mirrors in the central area, to allow those interested only in sneakers to see our attractive outfits.
- **Clustering** - KMeans, Hierarchical & DBSCAN algorithms each identified clusters that represented certain kinds of customers based on their purchasing activity. Each method had its strengths and weaknesses:
    - Hierarchical - best at identifying outliers
    - DBSCAN - best at subdividing the dense customer majority group (inliers)
    - KMeans - moderate quality and precision in both tasks, but generates the most balanced clusterings with balanced performance for both inliers and outliers
- Each of these algorithms identified very similar kinds of customer clusters - Bargain Hunters, Big Spenders, Collaboration Hunters... - indicating these are significant and reliable customer segments that exist in Six Squad's loyalty program, making them fit for use in developing personalized marketing and sales strategies.

NLP Data:

According to the word cloud, we can observe that 'customer service', 'experience', 'price', and 'staff' have a comparatively large size, meaning they are mentioned many times by reviewers. In terms of products, 'Jordan' seems to be a popular choice that Six Squad should stick to. The sentiment polarity distribution reveals a notably higher prevalence of neutral and positive comments, contrasting with fewer negative ones. This provides evidence that the selected stores maintain an excellent quality standard.

To better understand the nature of the review from different aspects, we performed Aspect-Based Sentiment Analysis. To reduce bias on aspect selection, we run a LDA model first to help us extract topics. We iterate the analysis over a range of potential topic numbers to build the LDA

model and assess each model's coherence score to measure its interpretability (meaningful separation of topics) and the optimal number of topics equals 5. The repeated emergence of terms related to "service" and "price" across different topic configurations highlights the significant emphasis customers place on these aspects when reviewing sneaker shops. After inputting the aspects derived from LDA & external research, we used the *BERT* model to classify each review as positive or negative. The results (see Figure 3) reveal a significant positive sentiment towards service (2409 positive mentions VS. 201 negative), suggesting customers generally appreciate the service quality in these sneaker shops. Product and brand aspects also receive predominantly positive feedback, indicating strong product satisfaction and brand perception. Interestingly, the price aspect shows a balanced view but leans towards positivity, suggesting customers find value in the products despite the cost. Quality and refund policies have fewer mentions but exhibit a positive trend, highlighting areas of strength while also indicating potential areas for further improvement or focus.

Also, we identify the top-performing sneaker stores across various customer satisfaction aspects, including service, price, refund policies, brand, quality, and product, by calculating the average sentiment scores for each aspect. It underscores the multifaceted nature of customer satisfaction and is beneficial for the SixSquad to find the best stores in each aspect to do shadowing.

5. Limitations & Future Improvements

Transaction Data: We have the following plans to take the analysis deeper -

- **MBA** - We plan to investigate the impact of discounts on customer behavior and their effect on increasing store revenue. Also, we can see how association rules change at different price levels - luxury items, moderate-cost items & low-cost items.
- **Clustering** - Further hyperparameter optimization may be done for the three methods to improve cluster quality and interpretability, particularly for DBSCAN. With the many categorical variables in the dataset, we can subdivide our clustering by store, gender , preferred Brand, Product category etc. There is potential for more targeted clustering.

NLP Data: Given the manual extraction of text data, the total quantity we can acquire is somewhat limited. Consequently, the number of significant aspects that LDA can capture is less than initially anticipated. With more time at our disposal, we would consider scraping a larger quantity of comments, relying solely on the aspects identified by LDA without additional manual supplementation. Additionally, we haven't utilized information on whether a reviewer is a local guide and their review/photo history. A local guide and a reviewer who has uploaded numerous reviews and photos can be considered more trustworthy. We might assign higher weights to reviews from such individuals and assess how our results might change.

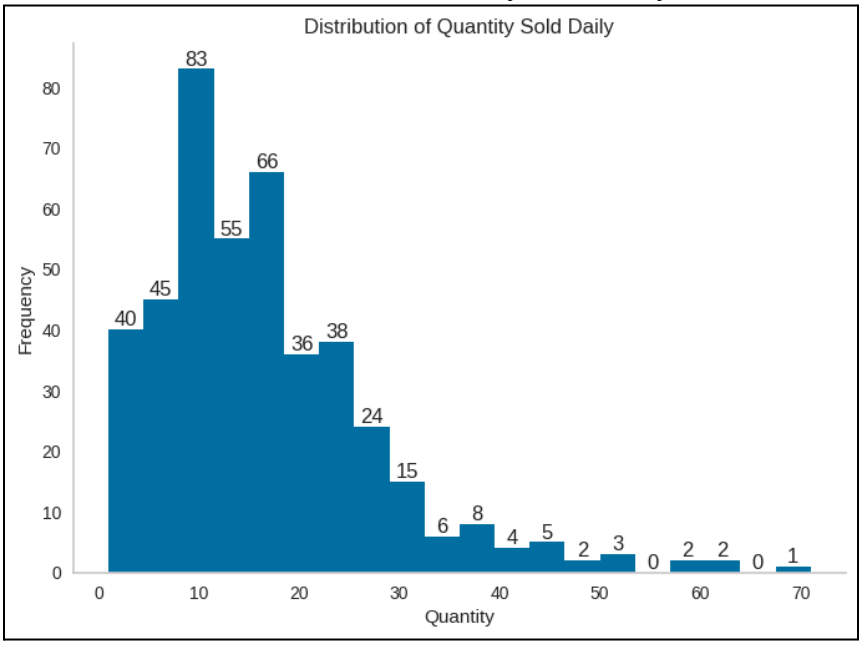
Link to Contribution Sheet: BA820 - Team 6
Link to GitHub Project: https://github.com/users/d-roho/projects/2

Project Deliverable 3 - Team A6

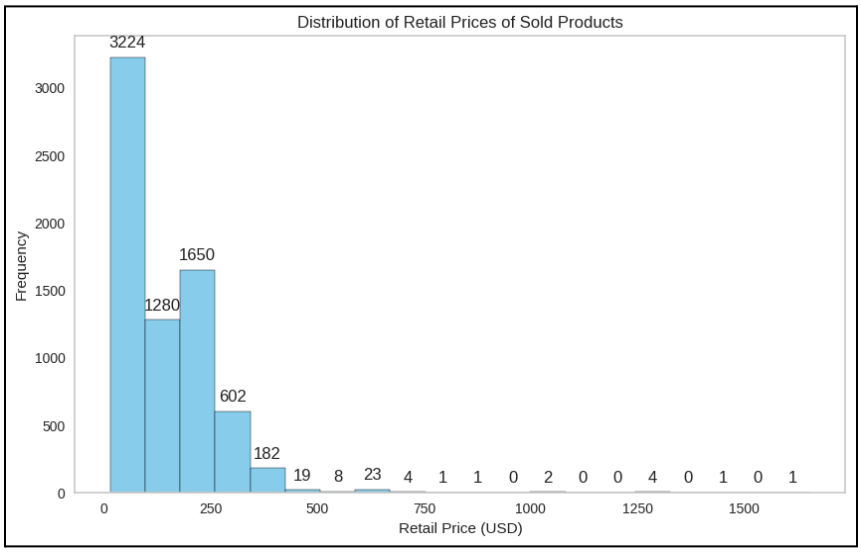
Appendix

Transaction Data EDA

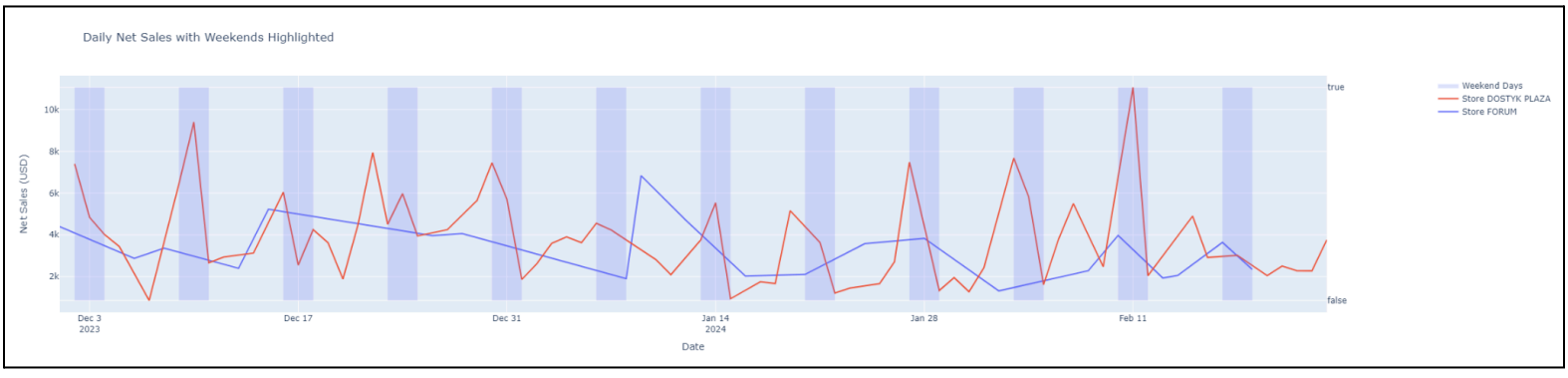
Distribution of Quantity Sold Daily



Distribution of Retail Prices of Sold Products



Daily Net Sales with Weekends Visualization



Project Deliverable 3 - Team A6

Association Rules

Top Five Frequent Itemsets

	support	itemsets
251	0.022662	(WMNS AIR JORDAN 1 LOW)
33	0.020022	(Bearbrick Blindbox series 46 by Medicom Toy)
42	0.019142	(CR7 Fashion, 2-Pack Trunk Mesh)
44	0.018482	(CR7 Trunk, 3-pack)
227	0.015842	(U J ED CUSH POLY CREW 3PR 144)

Top Five Associations Rules

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage
0	(Medicom Toy Bearbrick Blindbox series 45 by M...	(Bearbrick Blindbox series 46 by Medicom Toy)	0.015402	0.020022	0.005501	0.357143	17.837520	0.005192
1	(Bearbrick Blindbox series 46 by Medicom Toy)	(Medicom Toy Bearbrick Blindbox series 45 by M...	0.020022	0.015402	0.005501	0.274725	17.837520	0.005192
2	(CR7 Boys Trunk 2-pack)	(CR7 Kids Socks 3-pack)	0.014961	0.009461	0.003740	0.250000	26.424419	0.003599
3	(CR7 Kids Socks 3-pack)	(CR7 Boys Trunk 2-pack)	0.009461	0.014961	0.003740	0.395349	26.424419	0.003599
4	(CR7 Fashion Trunk Org 2-pack)	(CR7 Trunk, 3-pack)	0.012981	0.018482	0.001980	0.152542	8.253632	0.001740
5	(CR7 Trunk, 3-pack)	(CR7 Fashion Trunk Org 2-pack)	0.018482	0.012981	0.001980	0.107143	8.253632	0.001740

Top Five Frequent Itemsets in Weekdays

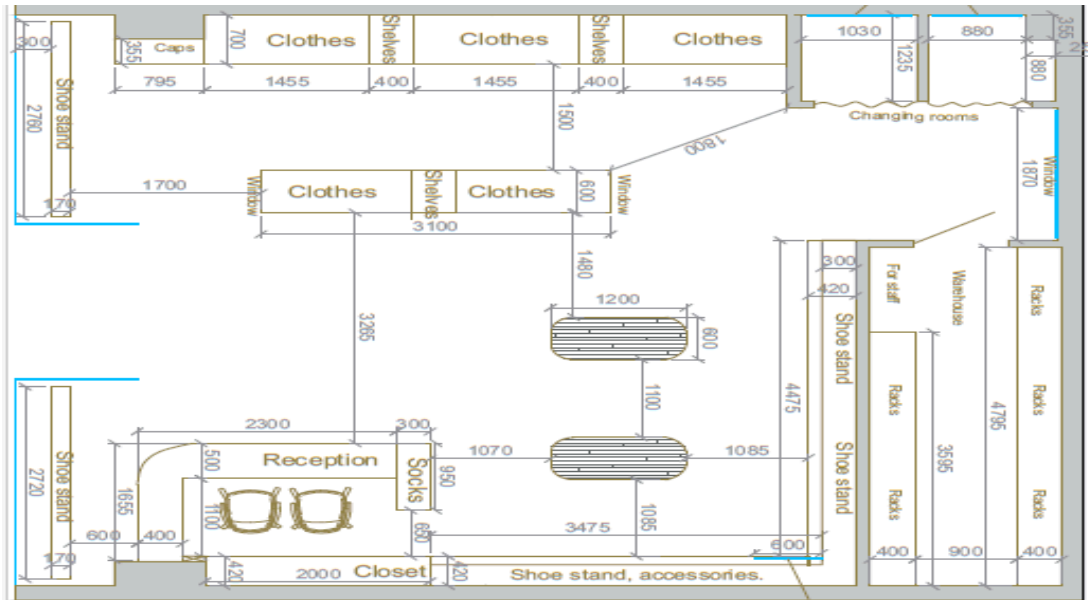
	support	itemsets
449	0.022026	(WMNS AIR JORDAN 1 LOW)
58	0.021659	(CR7 Fashion, 2-Pack Trunk Mesh)
60	0.020925	(CR7 Trunk, 3-pack)
415	0.018355	(U J ED CUSH POLY CREW 3PR 144)
48	0.015786	(Bearbrick Blindbox series 46 by Medicom Toy)

Top Five Frequent Itemsets in Weekdays

	support	itemsets
55	0.026359	(Bearbrick Blindbox series 46 by Medicom Toy)
452	0.023613	(WMNS AIR JORDAN 1 LOW)
64	0.021966	(CR7 Boys Trunk 2-pack)
274	0.020319	(Medicom Toy Bearbrick Blindbox series 45 by M...
138	0.018671	(Jordan Everyday Max_WHITE/WHITE/WHITE/BLACK)

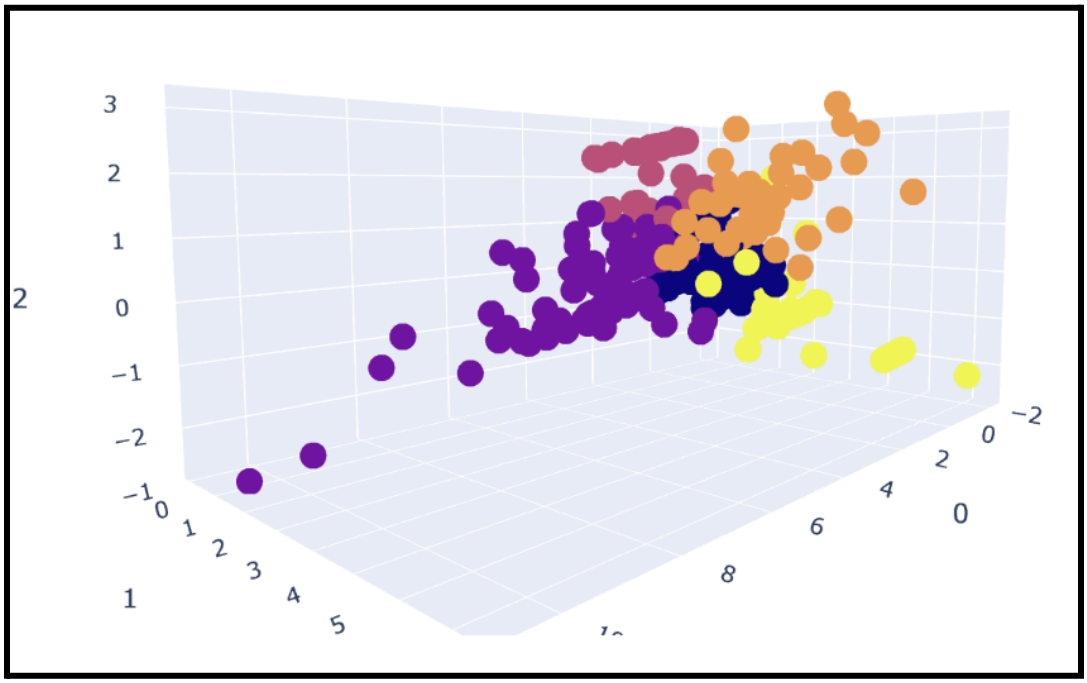
Project Deliverable 3 - Team A6

Proposed Store Layout



Clustering

KMeans Clusters Visualized using PCA (K=5)



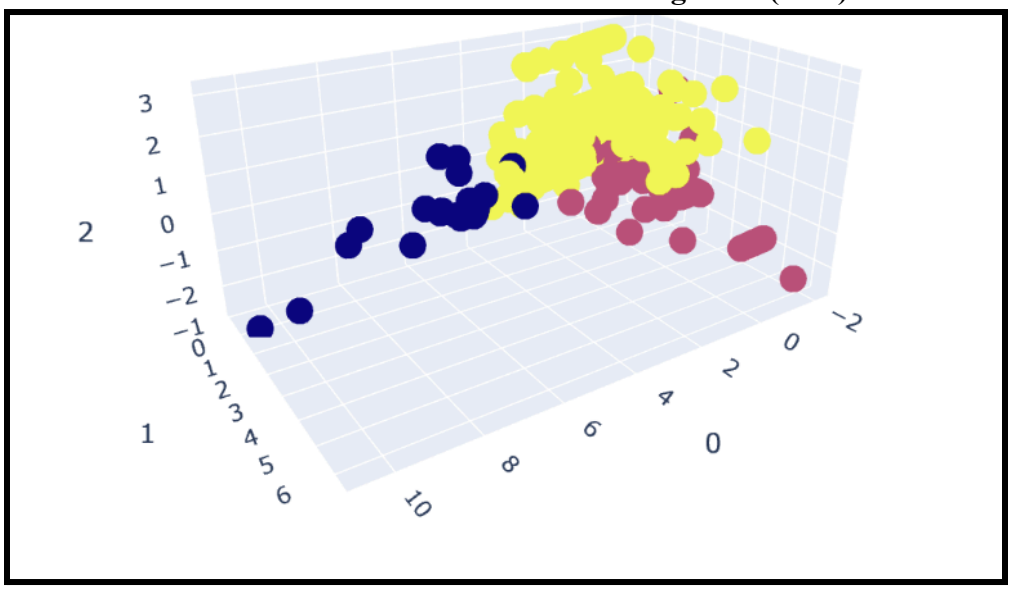
Project Deliverable 3 - Team A6

Average values of KMeans Clusters and Interpretation						
Cluster	**Total Transactions**	**Average Quantity**	**Average Discount Rate**	**Average Cart Value**	**Collaboration Rate**	**Size**
0	**1.238519**	**1.221728**	**0.008079**	**85467.28**	**0.010346**	**675**
1	**1.279279**	**3.509009**	**0.018489**	**252748.2**	**0.123244**	**111**
2	**1.169811**	**1.378931**	**0.006773**	**103445.3**	**0.816262**	**106**
3	**5.380952**	**1.890788**	**0.053206**	**114442.9**	**0.177192**	**42**
4	**1.125**	**1.2875**	**0.664346**	**64111.94**	**0.1**	**40**

Interpretation of K-Means Clustering

- **Cluster 0: Collaboration Hunters - These customers only purchase items that are collaboration, and pay full price.**
- **Cluster 1: Occasional Buyers - These are the bulk of customers who shop occasionally, and purchase relatively low cost items at full price**
- **Cluster 2: The Regulars - These are regular customers that purchase medium priced items very often, at a moderate discount. A moderate preference for collabs**
- **Cluster 3: Bargain Hunters - Only buy low cost items at a heavy discount**
- **Cluster 4: The Whales - Customers that shop occasionally but spend heavily each visit.**

Hierarchical Clusters Visualized using PCA (K=3)



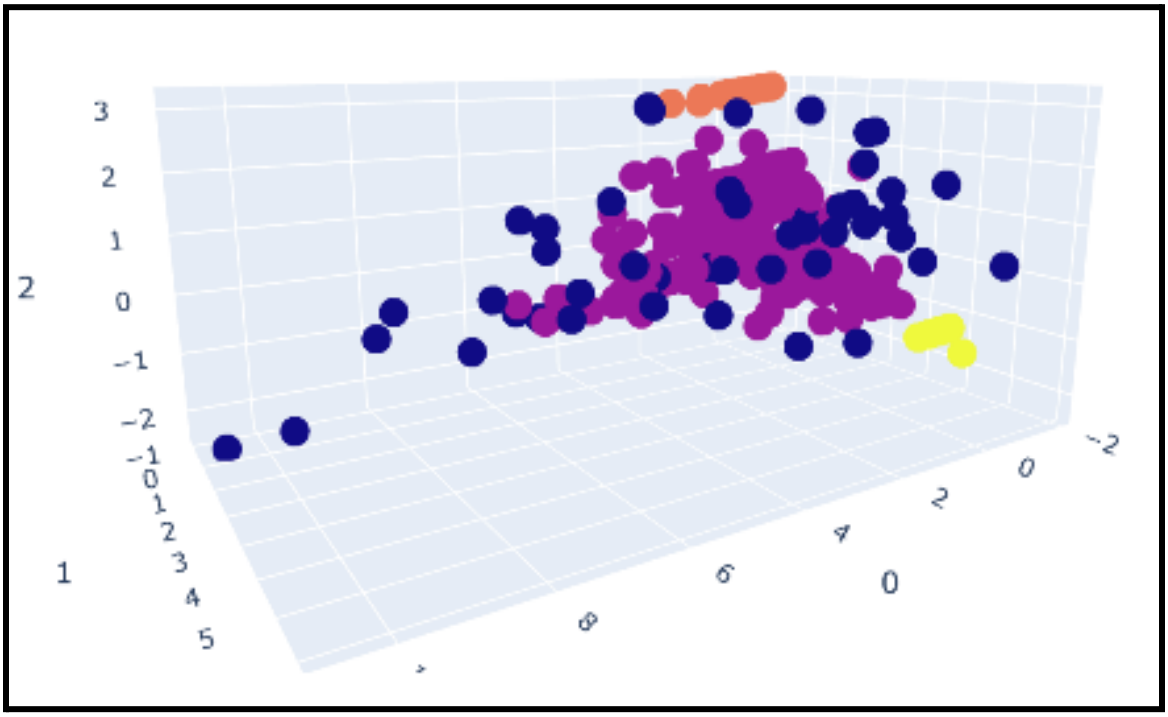
Project Deliverable 3 - Team A6

Average values of Hierarchical Clusters and Interpretation						
Cluster	**Total Transactions**	**Average Quantity**	**Average Discount Rate**	**Average Cart Value**	**Collaboration Rate**	**Size**
1	**1.333333**	**5.460317**	**0.020818**	**413296.4**	**0.156665**	**21**
2	**1.224138**	**1.5**	**0.528846**	**81434.7**	**0.096839**	**58**
3	**1.423464**	**1.440871**	**0.006616**	**101317.7**	**0.122595**	**895**

Interpretation of Hierarchical Clustering

- **Cluster 1: The Whales - Customers that shop occasionally but spend heavily each visit.**
- **Cluster 2: Bargain Hunters - Only buy low cost items at a heavy discount**
- **Cluster 3: The Regulars - customers that purchase medium priced items occasionally, at full price.**

DBSCAN Clusters Visualized (K=4)



Project Deliverable 3 - Team A6

NLP Section

Figure 2.1 Word Cloud for All the Reviews

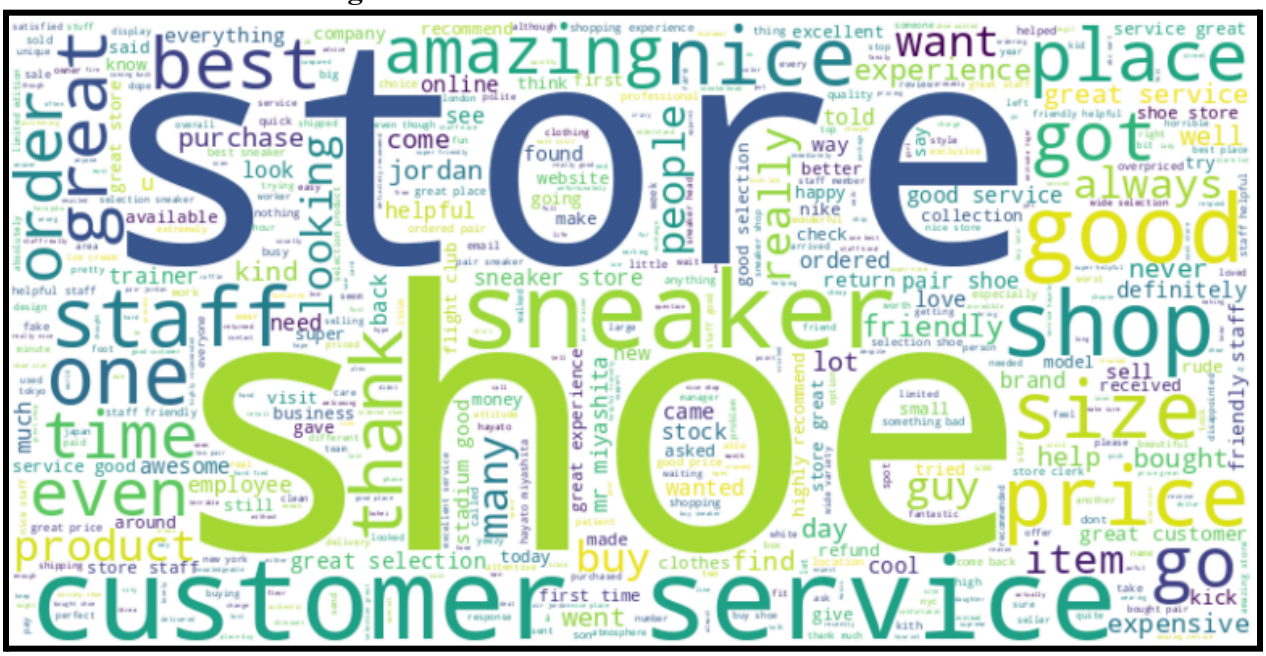
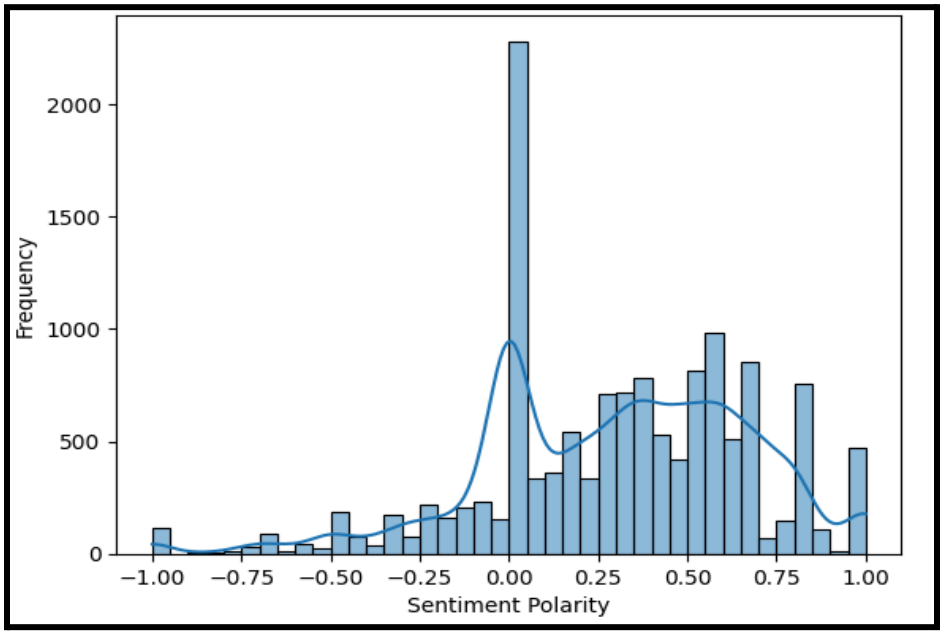
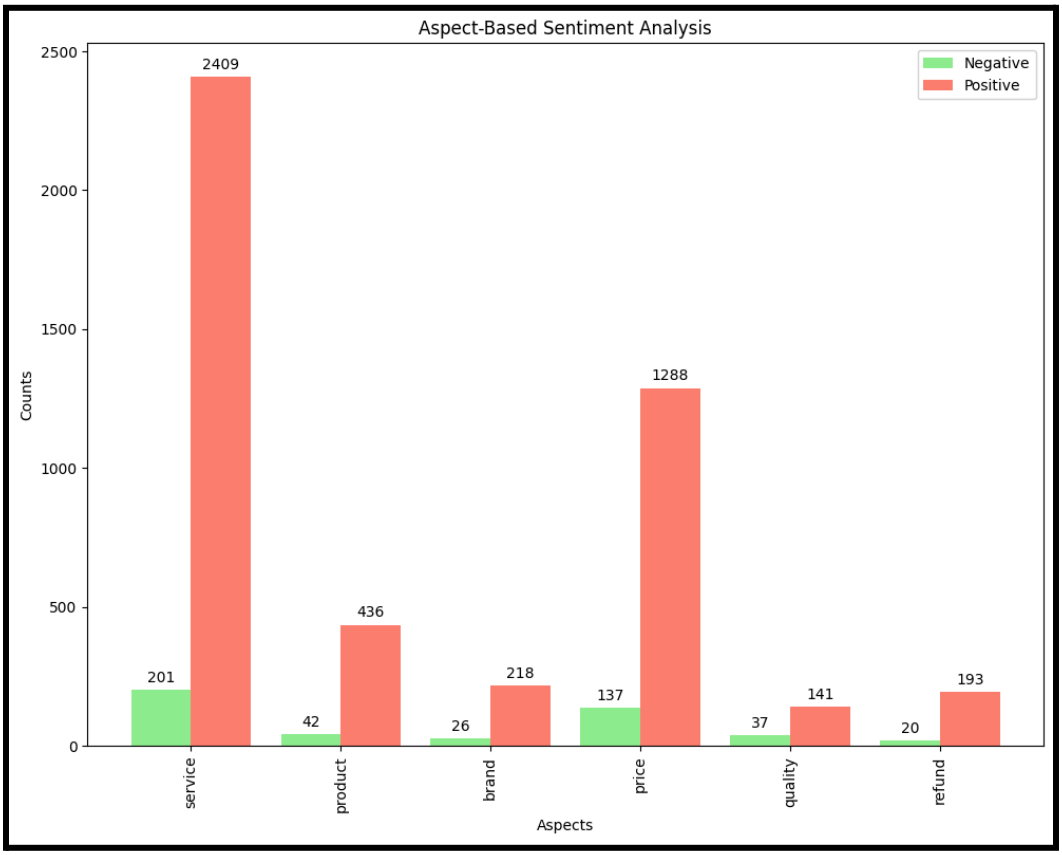


Figure 2.2 Distribution of Sentiment Polarity



Project Deliverable 3 - Team A6

Figure 2.3 ASBA Results



Sentiment score for each STORE: taking Kicks Lab. as an example

```
Store: Kicks Lab.
 - Aspect: service, Average Sentiment: Positive (0.90)
 - Aspect: product, Average Sentiment: Positive (0.91)
 - Aspect: price, Average Sentiment: Positive (0.88)
 - Aspect: quality, Average Sentiment: Positive (0.70)
 - Aspect: brand, Average Sentiment: Positive (0.80)
 - Aspect: refund, Average Sentiment: Negative (0.40)
```

Sentiment score for each store in different ASPECTS: taking 'price' as an example

```
Aspect: price
  1. Store: District One, Average Sentiment Score: 0.97
  2. Store: Origins NYC, Average Sentiment Score: 0.94
  3. Store: Presented By, Average Sentiment Score: 0.93
```