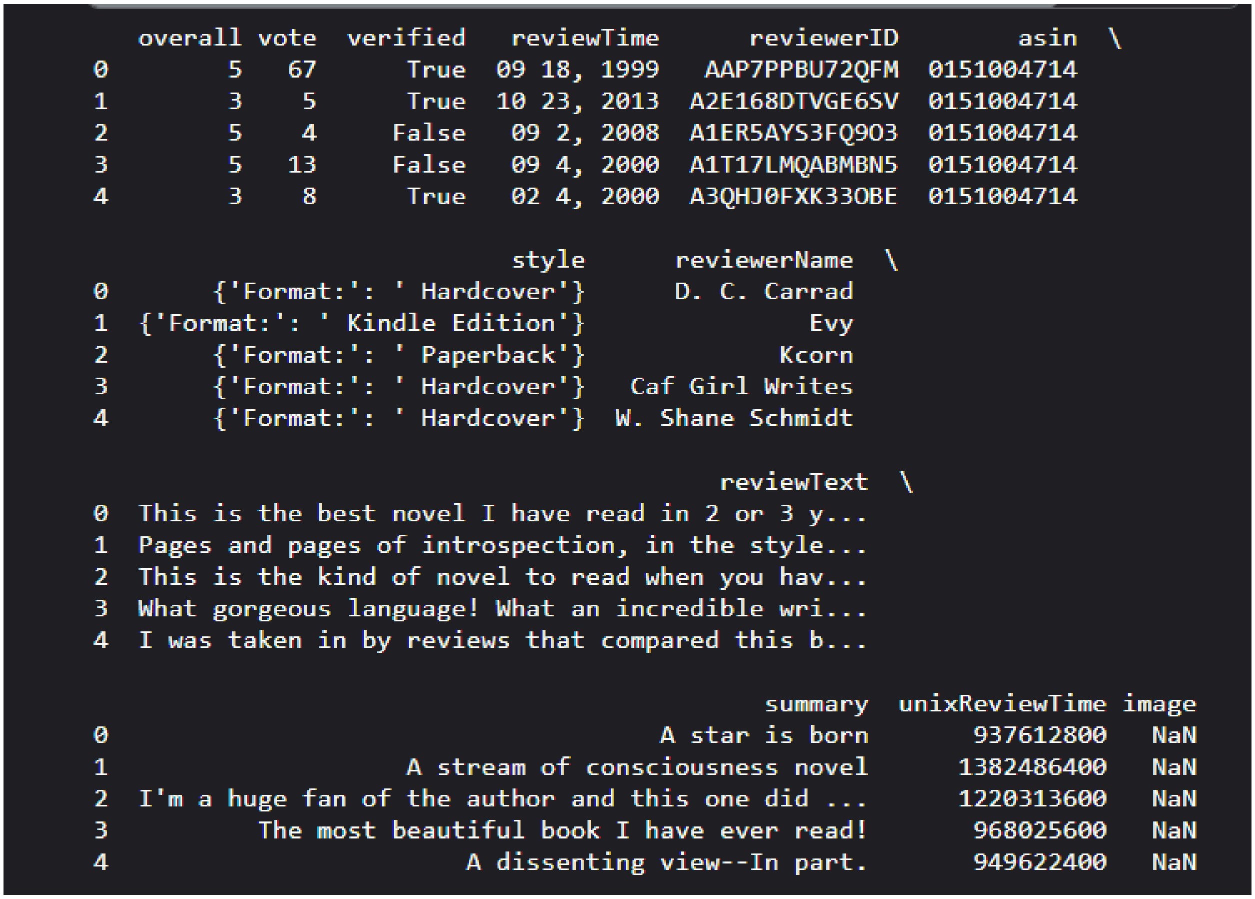
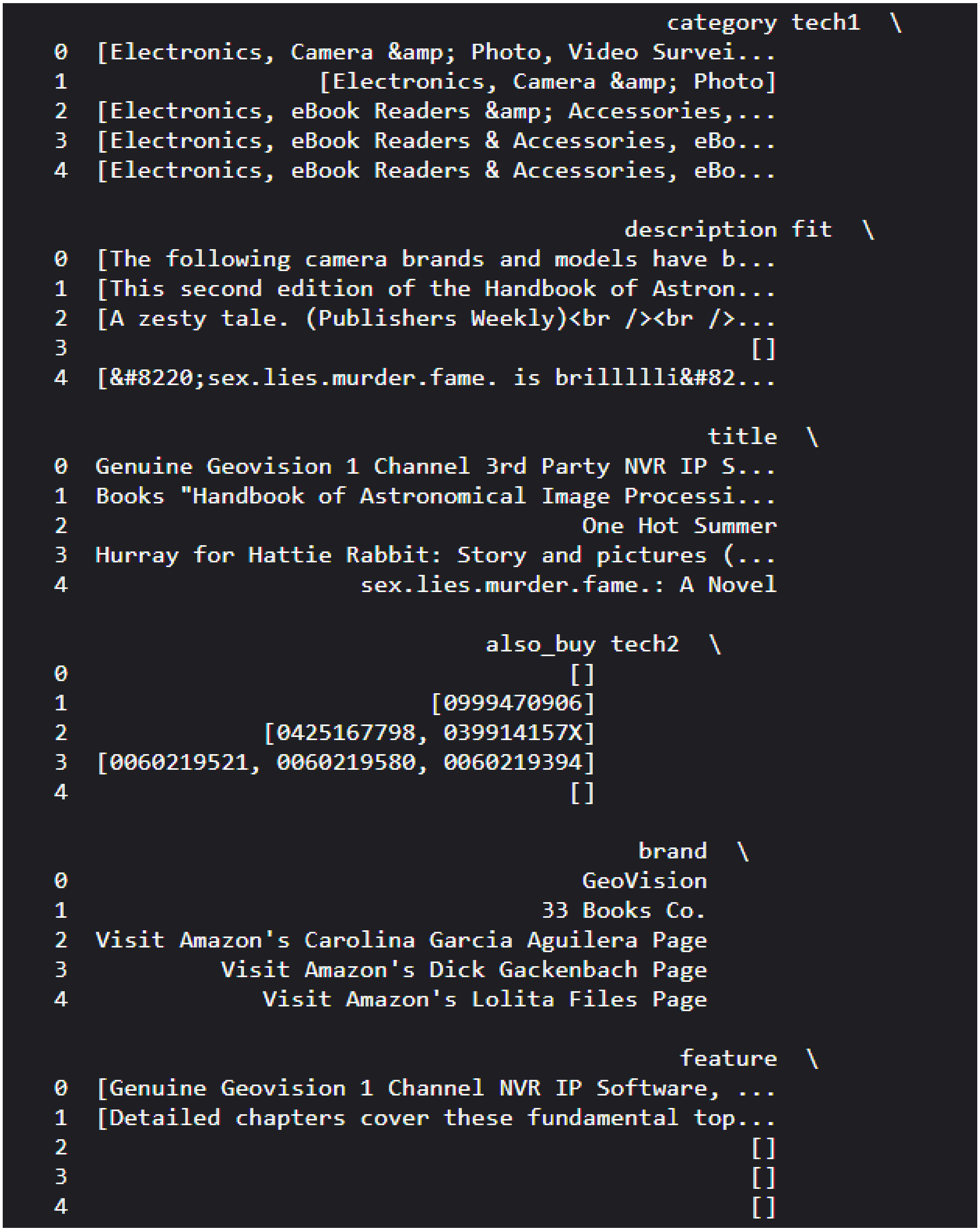
IR report

Reading and Preprocessing Data:

- Loaded the Amazon Reviews dataset and the metadata dataset for Electronics category into dataframes using Pandas.
- Filtered the metadata dataframe to select only products related to 'Headphones'.
- Handled missing values by filling them with a placeholder ('Missing') and removed duplicate entries.



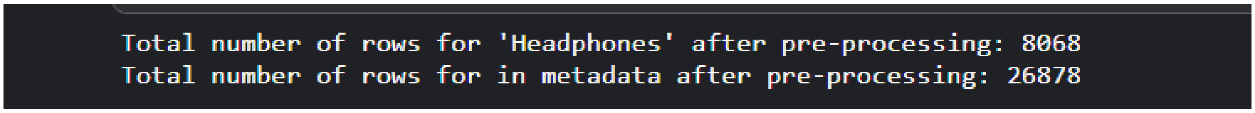
```
   overall vote  verified   reviewTime      reviewerID        asin  \
0        5   67      True  09 18, 1999   AAP7PPBU72QFM  0151004714   
1        3    5      True  10 23, 2013  A2E168DTVGE6SV  0151004714   
2        5    4     False   09 2, 2008  A1ER5AYS3FQ9O3  0151004714   
3        5   13     False   09 4, 2000  A1T17LMQABMBN5  0151004714   
4        3    8      True   02 4, 2000  A3QHJ0FXK33OBE  0151004714   

                            style      reviewerName  \
0      {'Format:': ' Hardcover'}      D. C. Carrad   
1  {'Format:': ' Kindle Edition'}               Evy   
2      {'Format:': ' Paperback'}             Kcorn   
3      {'Format:': ' Hardcover'}   Caf Girl Writes   
4      {'Format:': ' Hardcover'}  W. Shane Schmidt   

                                        reviewText  \
0  This is the best novel I have read in 2 or 3 y...   
1  Pages and pages of introspection, in the style...   
2  This is the kind of novel to read when you hav...   
3  What gorgeous language! What an incredible wri...   
4  I was taken in by reviews that compared this b...   

                                           summary  unixReviewTime image  
0                                   A star is born       937612800   NaN  
1                  A stream of consciousness novel      1382486400   NaN  
2  I'm a huge fan of the author and this one did ...      1220313600   NaN  
3          The most beautiful book I have ever read!       968025600   NaN  
4                      A dissenting view--In part.       949622400   NaN  
```



```
                                            category tech1  \
0  [Electronics, Camera &amp; Photo, Video Survei...
1                  [Electronics, Camera &amp; Photo]
2  [Electronics, eBook Readers &amp; Accessories,...
3  [Electronics, eBook Readers & Accessories, eBo...
4  [Electronics, eBook Readers & Accessories, eBo...

                                         description fit  \
0  [The following camera brands and models have b...
1  [This second edition of the Handbook of Astron...
2  [A zesty tale. (Publishers Weekly)<br /><br />...
3                                                 []
4  [&#8220;sex.lies.murder.fame. is brillllli&#82...

                                               title  \
0  Genuine Geovision 1 Channel 3rd Party NVR IP S...
1  Books "Handbook of Astronomical Image Processi...
2                                     One Hot Summer
3  Hurray for Hattie Rabbit: Story and pictures (...
4                     sex.lies.murder.fame.: A Novel

                               also_buy tech2  \
0                                    []
1                          [0999470906]
2              [0425167798, 039914157X]
3  [0060219521, 0060219580, 0060219394]
4                                    []

                                          brand  \
0                                     GeoVision
1                                  33 Books Co.
2  Visit Amazon's Carolina Garcia Aguilera Page
3           Visit Amazon's Dick Gackenbach Page
4              Visit Amazon's Lolita Files Page

                                             feature  \
0  [Genuine Geovision 1 Channel NVR IP Software, ...
1  [Detailed chapters cover these fundamental top...
2                                                 []
3                                                 []
4                                                 []
```

Reporting Total Number of Rows:

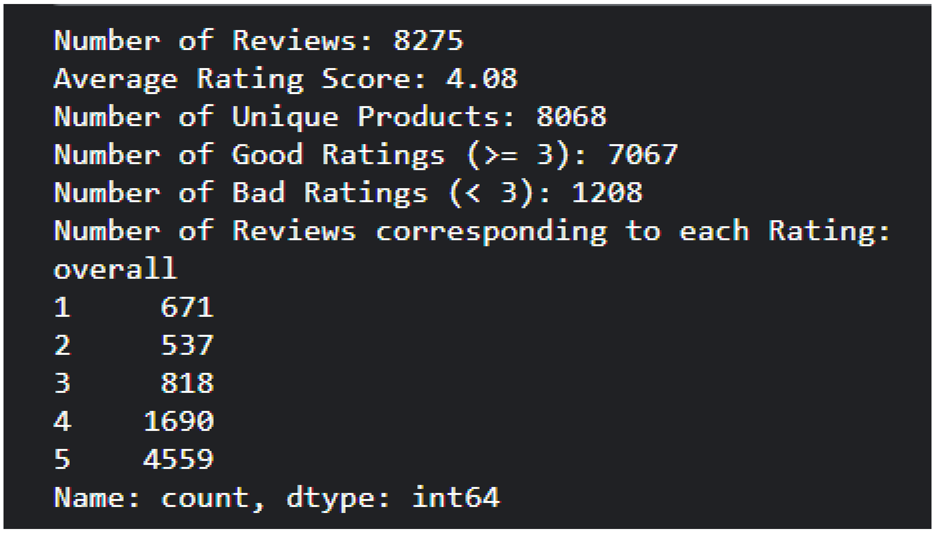
- Calculated the total number of rows in the preprocessed 'Headphones' dataframe.



```
Total number of rows for 'Headphones' after pre-processing: 8068
Total number of rows for in metadata after pre-processing: 26878
```

Descriptive Statistics:

- Calculated descriptive statistics such as number of reviews, average rating score, number of unique products, number of good and bad ratings based on a threshold, and number of reviews corresponding to each rating.



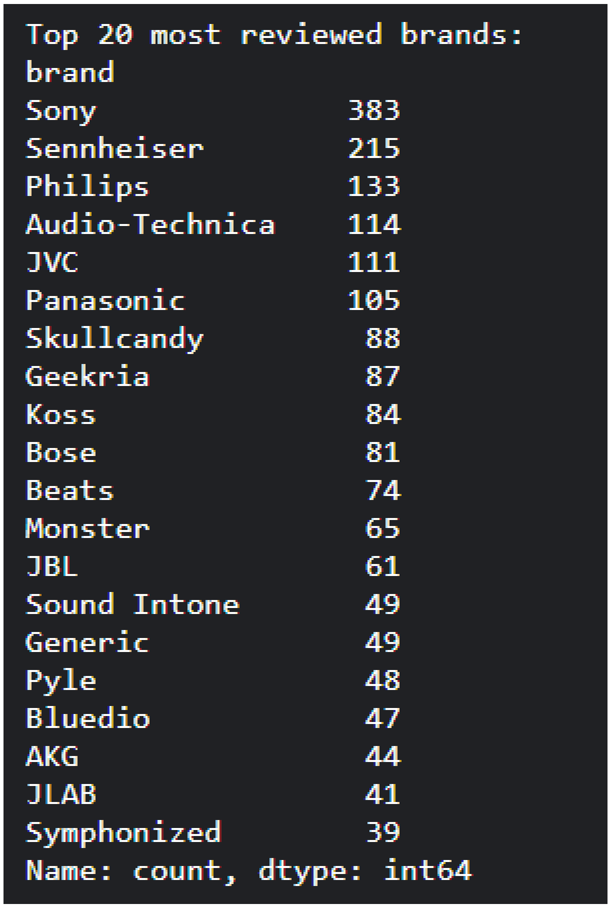
```
Number of Reviews: 8275
Average Rating Score: 4.08
Number of Unique Products: 8068
Number of Good Ratings (>= 3): 7067
Number of Bad Ratings (< 3): 1208
Number of Reviews corresponding to each Rating:
overall
1     671
2     537
3     818
4    1690
5    4559
Name: count, dtype: int64
```

Text Preprocessing:

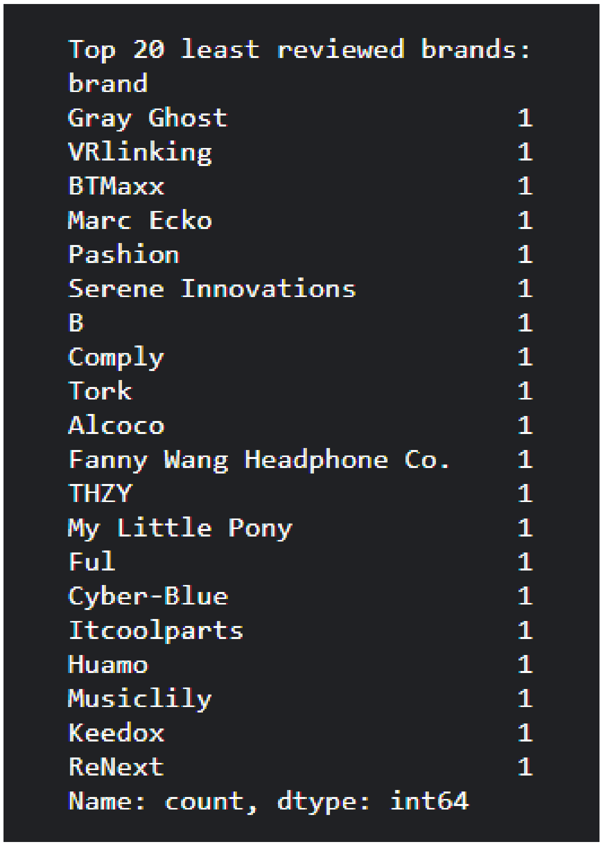
- Implemented a pipeline for text preprocessing, which involved removing HTML tags, accented characters, expanding acronyms, removing special characters, and lemmatization.

Exploratory Data Analysis (EDA):

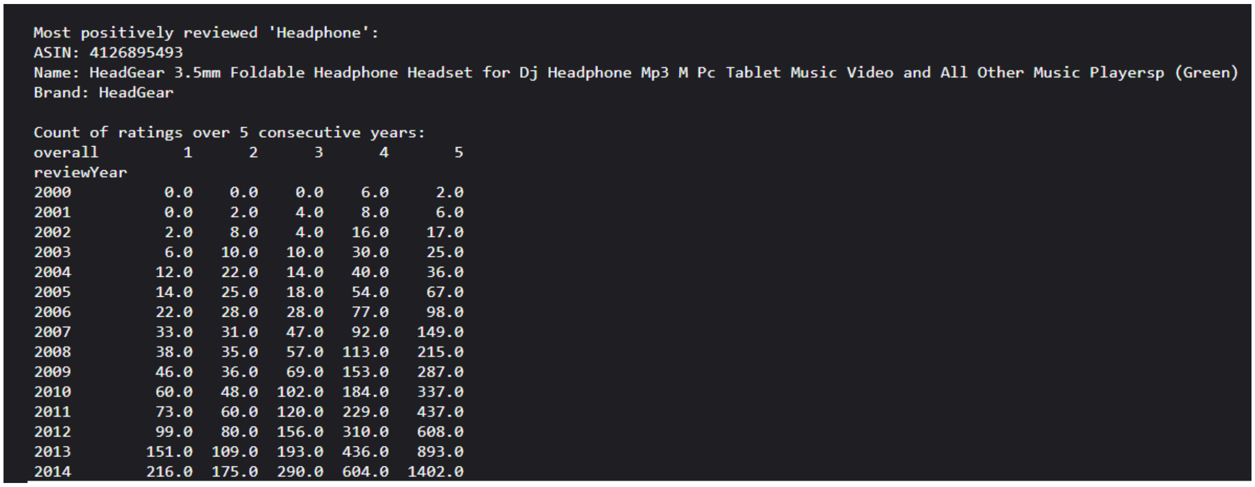
- Identified the top 20 most and least reviewed brands in the selected category.
- Determined the most positively reviewed headphone based on the highest overall rating.
- Analyzed the count of ratings over 5 consecutive years.
- Generated word clouds for 'Good' and 'Bad' reviews to visualize commonly used words.
- Plotted a pie chart showing the distribution of ratings vs. the number of reviews.



```
Top 20 most reviewed brands:
brand
Sony                383
Sennheiser          215
Philips             133
Audio-Technica      114
JVC                 111
Panasonic           105
Skullcandy           88
Geekria              87
Koss                 84
Bose                 81
Beats                74
Monster              65
JBL                  61
Sound Intone         49
Generic              49
Pyle                 48
Bluedio              47
AKG                  44
JLAB                 41
Symphonized          39
Name: count, dtype: int64
```

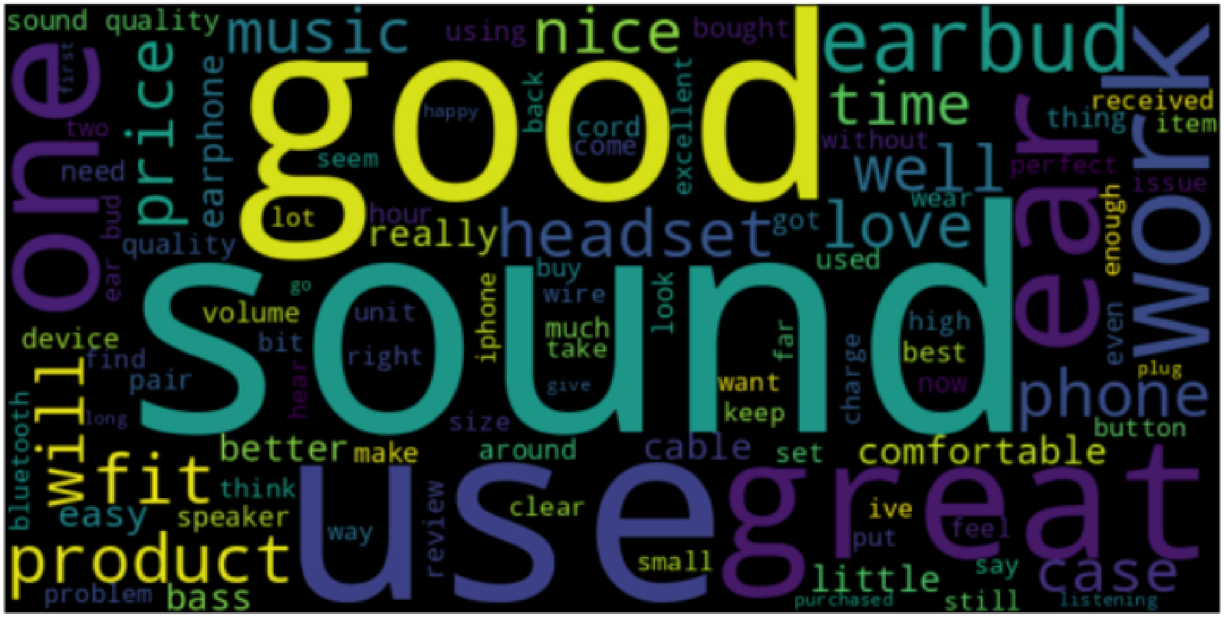


```
Top 20 least reviewed brands:
brand
Gray Ghost                     1
VRlinking                      1
BTMaxx                         1
Marc Ecko                      1
Pashion                        1
Serene Innovations             1
B                              1
Comply                         1
Tork                           1
Alcoco                         1
Fanny Wang Headphone Co.       1
THZY                           1
My Little Pony                 1
Ful                            1
Cyber-Blue                     1
Itcoolparts                    1
Huamo                          1
Musiclily                      1
Keedox                         1
ReNext                         1
Name: count, dtype: int64
```

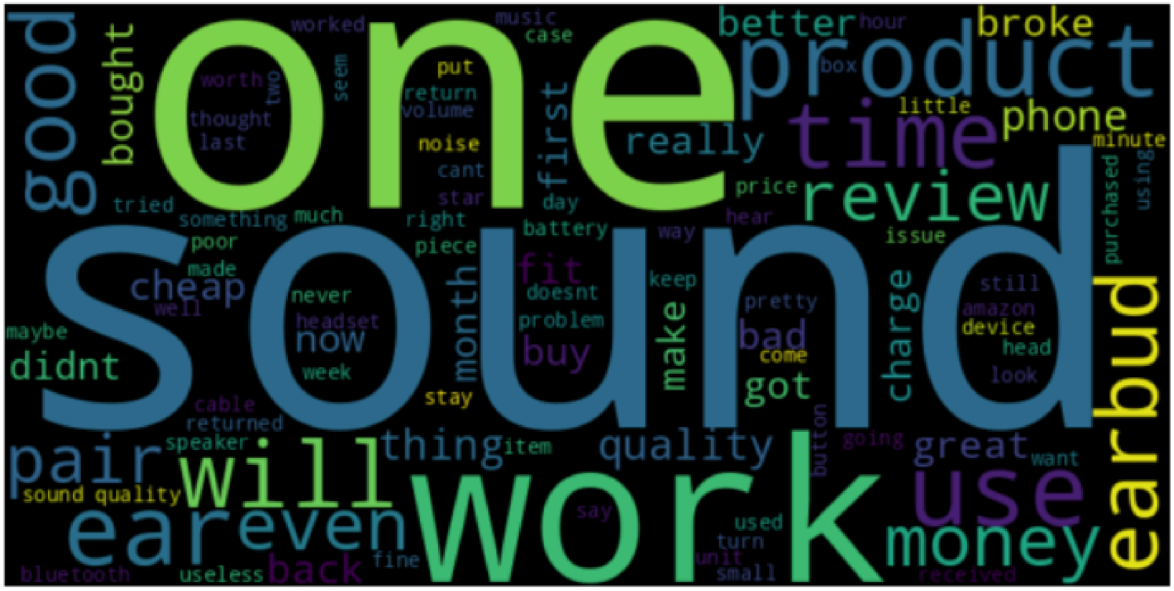


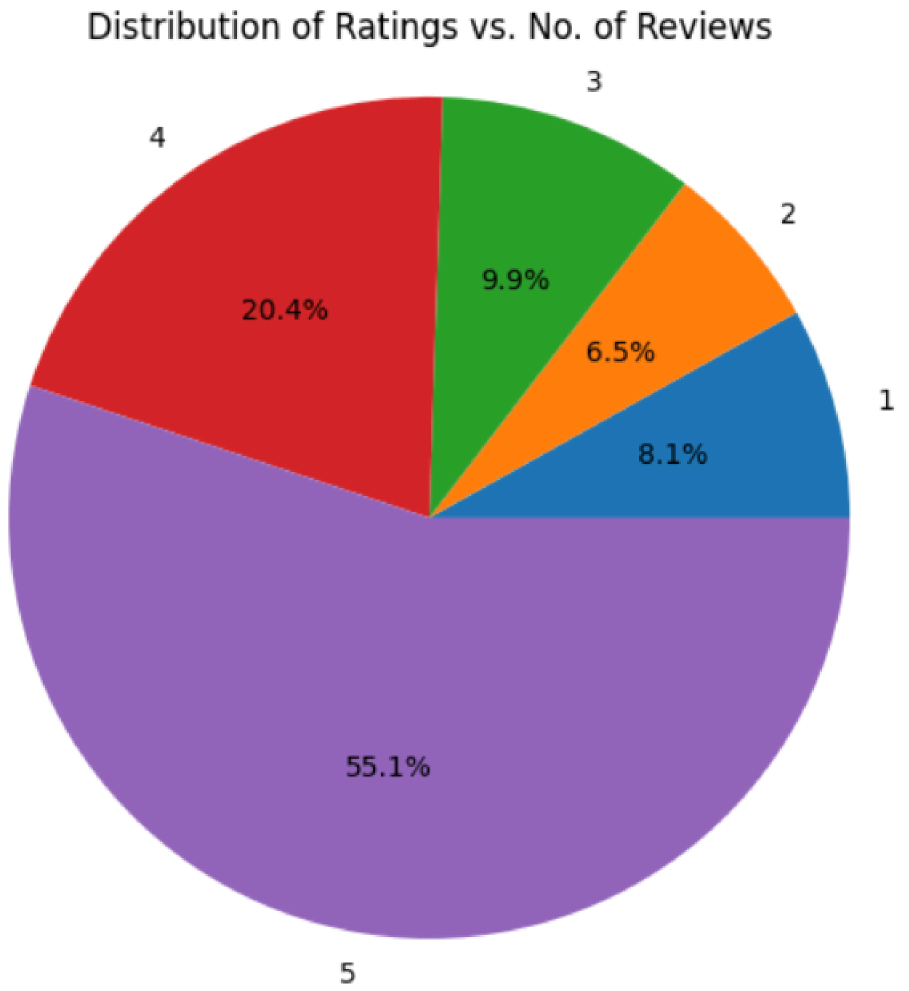
```
Most positively reviewed 'Headphone':
ASIN: 4126895493
Name: HeadGear 3.5mm Foldable Headphone Headset for Dj Headphone Mp3 M Pc Tablet Music Video and All Other Music Playersp (Green)
Brand: HeadGear

Count of ratings over 5 consecutive years:
overall          1      2      3      4       5
reviewYear
2000           0.0    0.0    0.0    6.0     2.0
2001           0.0    2.0    4.0    8.0     6.0
2002           2.0    8.0    4.0   16.0    17.0
2003           6.0   10.0   10.0   30.0    25.0
2004          12.0   22.0   14.0   40.0    36.0
2005          14.0   25.0   18.0   54.0    67.0
2006          22.0   28.0   28.0   77.0    98.0
2007          33.0   31.0   47.0   92.0   149.0
2008          38.0   35.0   57.0  113.0   215.0
2009          46.0   36.0   69.0  153.0   287.0
2010          60.0   48.0  102.0  184.0   337.0
2011          73.0   60.0  120.0  229.0   437.0
2012          99.0   80.0  156.0  310.0   608.0
2013         151.0  109.0  193.0  436.0   893.0
2014         216.0  175.0  290.0  604.0  1402.0
```

Good Reviews



Bad Reviews





Assumptions:

- The dataset provided covers a representative sample of 'Headphones' products in the Electronics category.
- Ratings are assumed to be reliable indicators of product quality.
- Text preprocessing techniques applied are sufficient for cleaning and normalizing review text data.
- The threshold for categorizing ratings as 'Good' or 'Bad' is set at >=3, assuming a neutral rating lies between 3 and 4.
- Brand names in the metadata are accurate and consistent.

Feature Engineering with Text Data:

- Utilized the Review Text as input feature and Rating Class as the target variable.

- Applied a relevant feature engineering technique, in this case, TF-IDF (Term Frequency-Inverse Document Frequency), to convert text data into numerical vectors while considering the importance of words in a document corpus.

Data Splitting

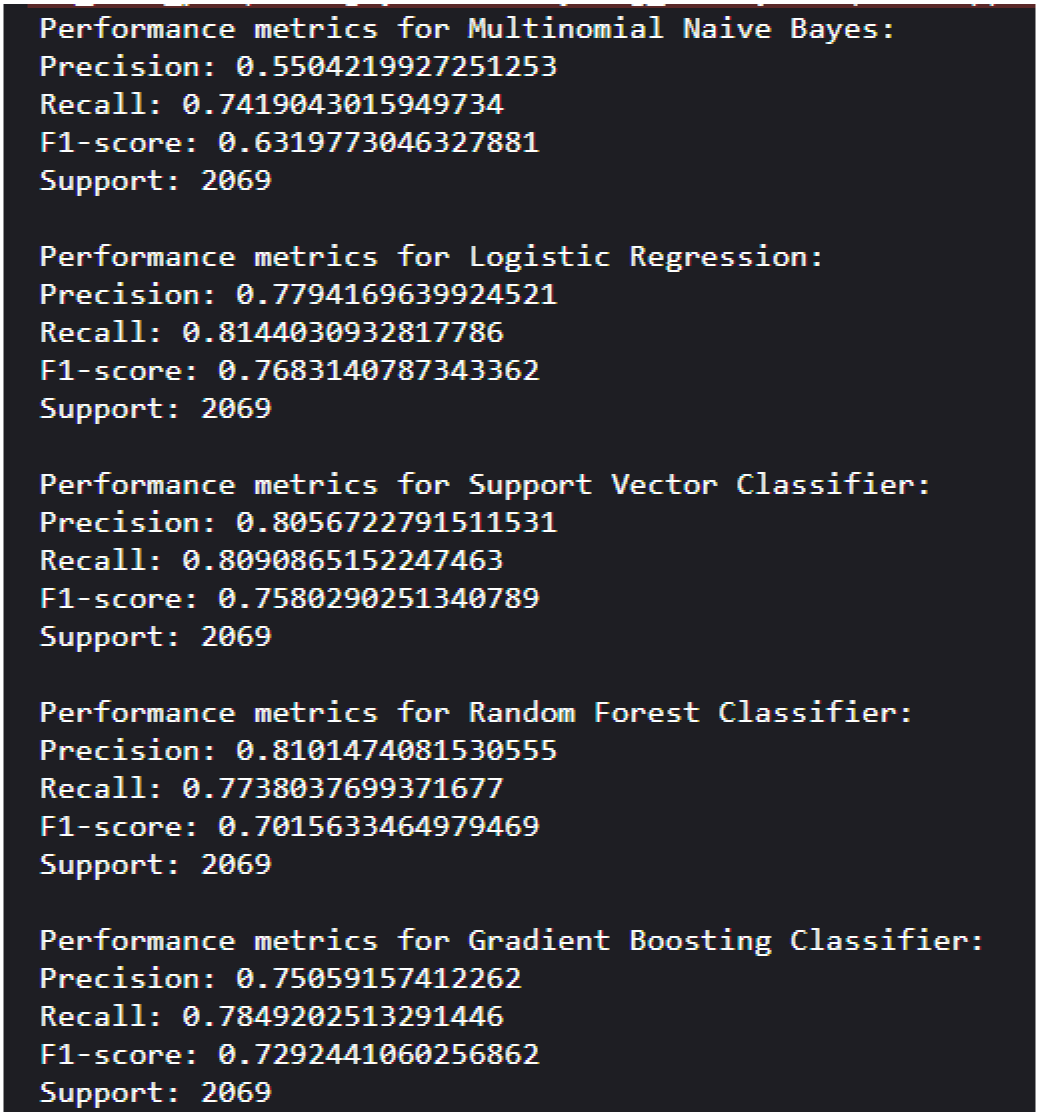
- Split the dataset into training and testing sets in a 75:25 ratio to evaluate model performance effectively.

Model Selection and Evaluation

- Selected 5 machine learning-based models: Multinomial Naive Bayes, Logistic Regression, Support Vector Classifier, Random Forest Classifier, and Gradient Boosting Classifier.

- Encoded the target variable (Rating Class) into three categories: 'Good', 'Average', and 'Bad'.

- Trained each model on the training set and evaluated their performance on the testing set using precision, recall, F1-score, and support metrics for each target class separately.



```
Performance metrics for Multinomial Naive Bayes:
Precision: 0.5504219927251253
Recall: 0.7419043015949734
F1-score: 0.6319773046327881
Support: 2069

Performance metrics for Logistic Regression:
Precision: 0.7794169639924521
Recall: 0.8144030932817786
F1-score: 0.7683140787343362
Support: 2069

Performance metrics for Support Vector Classifier:
Precision: 0.8056722791511531
Recall: 0.8090865152247463
F1-score: 0.7580290251340789
Support: 2069

Performance metrics for Random Forest Classifier:
Precision: 0.8101474081530555
Recall: 0.7738037699371677
F1-score: 0.7015633464979469
Support: 2069

Performance metrics for Gradient Boosting Classifier:
Precision: 0.75059157412262
Recall: 0.7849202513291446
F1-score: 0.7292441060256862
Support: 2069
```

Assumptions

- Rating Class is assumed to be an appropriate proxy for sentiment analysis, where ratings greater than 3 are considered 'Good', ratings equal to 3 are 'Average', and ratings less than 3 are 'Bad'.
- The TF-IDF vectorization technique is chosen based on its effectiveness in capturing word importance while considering document frequency.
- The dataset is assumed to be representative and balanced across different rating classes, ensuring fair evaluation of model performance.

User-Item Collaborative Filtering:
a) Creating User-Item Rating Matrix: First, we construct a matrix where each row represents a user, each column represents an item (product), and the entries denote the ratings given by users to items. This matrix is filled with zeros for missing ratings.
b) Normalizing Ratings: We normalize the ratings using Min-Max scaling, ensuring that ratings are scaled to a range between 0 and 1.
c) User-User Recommender System:

- Finding Similar Users: We calculate cosine similarity between users based on their ratings. For each user, we find the top N most similar users.
- K-Folds Validation: We divide the data into K folds and iterate over them. For each fold, one subset is used for validation, and the remaining subsets are used for training.
- Predicting Missing Values: We predict the missing ratings in the validation set using the training set and calculate the error between actual and predicted ratings.
- Reporting MAE: We calculate the Mean Absolute Error (MAE) for different values of N (number of similar users).

Item-Item Collaborative Filtering:
a) Creating Item-User Rating Matrix: Similar to the user-item matrix, we construct a matrix where each row represents an item and each column represents a user. Entries denote the ratings given by users to items.
b) Normalizing Ratings: We again normalize the ratings using Min-Max scaling.
c) Item-Item Recommender System:

- Finding Similar Items: We calculate cosine similarity between items based on user ratings. For each item, we find the top N most similar items.
- K-Folds Validation: Similar to user-user recommender, we use K-Folds validation to evaluate the performance.
- Predicting Missing Values: We predict the missing ratings in the validation set using the training set and calculate the error.
- Reporting MAE: We calculate the MAE for different values of N.

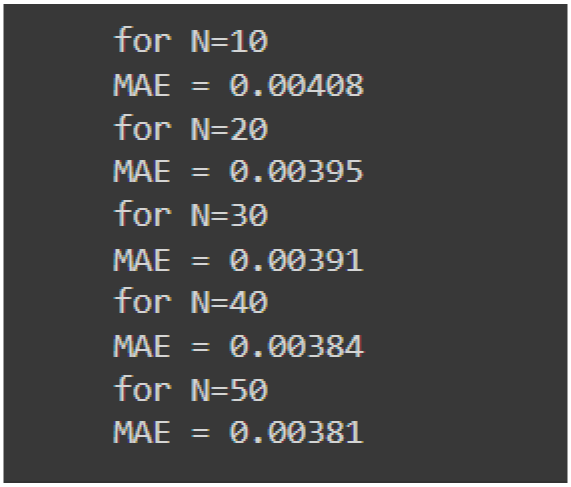
Assumptions:

- We assume that the ratings provided by users are meaningful indicators of their preferences and that similar users or items tend to have similar ratings.
- We assume that the datasets contain sufficient data for meaningful similarity calculations and predictions.
- We assume that Min-Max scaling is appropriate for normalizing ratings, providing a consistent range for similarity calculations.

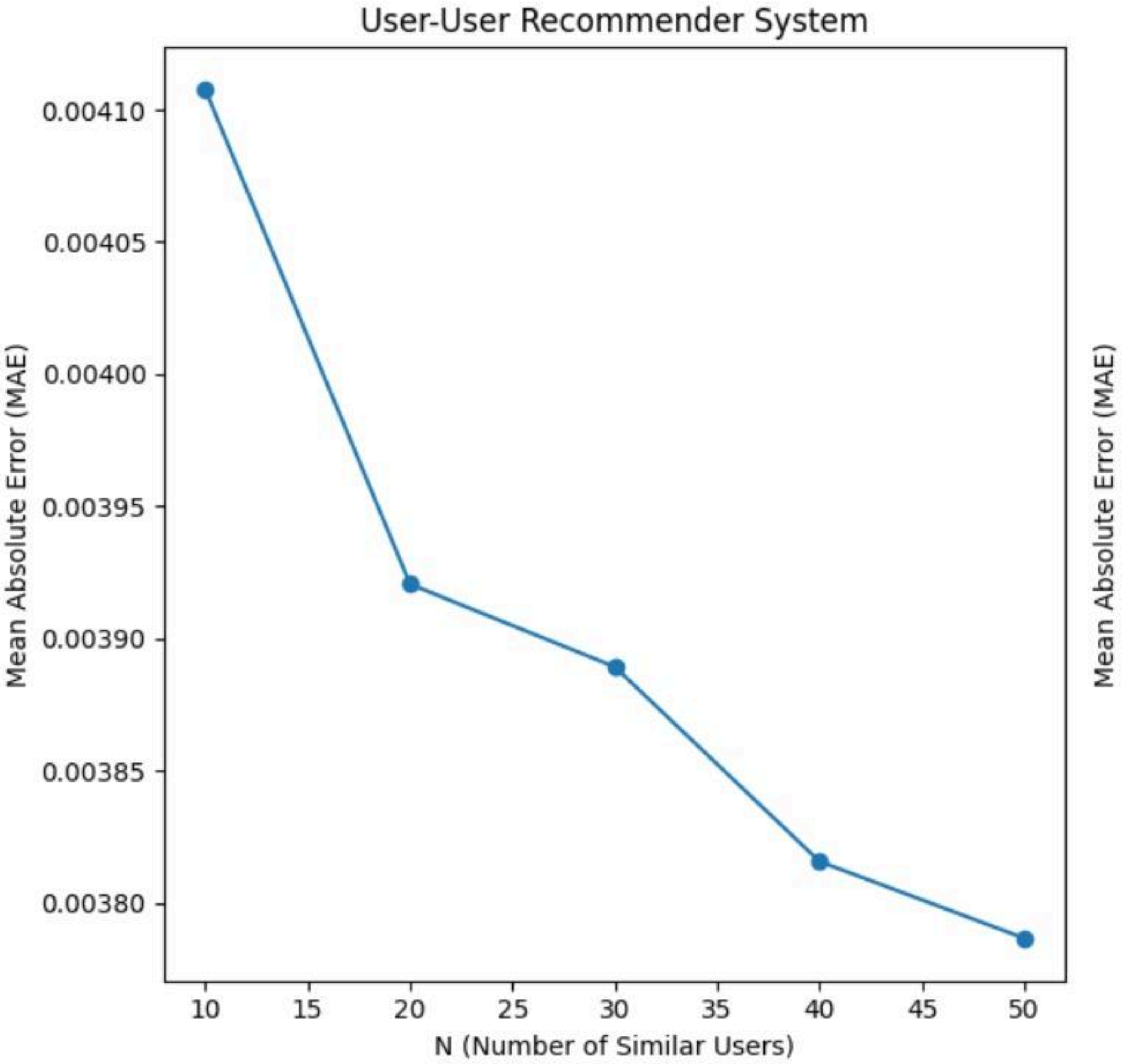
Methodologies:

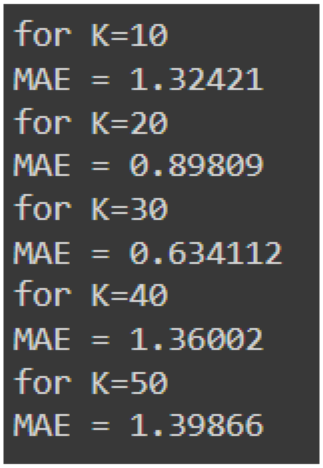
- We use cosine similarity for finding similar users and items, as it measures the cosine of the angle between two vectors of ratings.
- We employ K-Folds cross-validation to evaluate the performance of the recommender systems and ensure robustness of the models.
- We calculate the Mean Absolute Error (MAE) as a metric to quantify the accuracy of the predictions, providing insight into how well the recommender systems perform in predicting missing ratings.

Overall, these approaches enable us to build and evaluate collaborative filtering-based recommender systems for both user-item and item-item recommendations, providing personalized recommendations based on user preferences and item similarities.

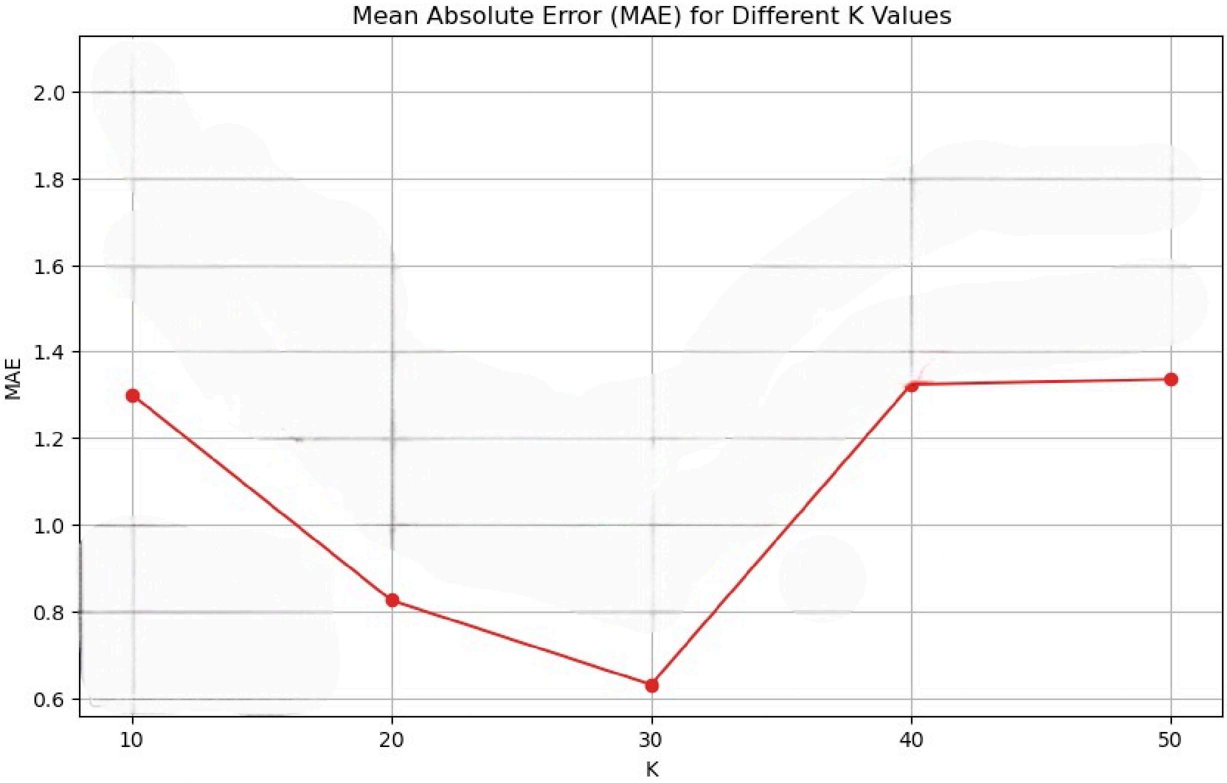


```
for N=10
MAE = 0.00408
for N=20
MAE = 0.00395
for N=30
MAE = 0.00391
for N=40
MAE = 0.00384
for N=50
MAE = 0.00381
```

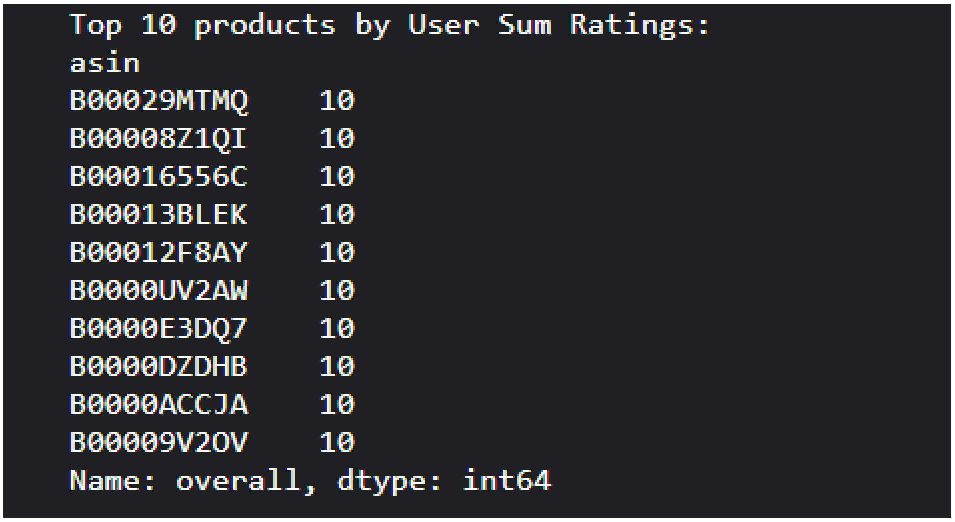

User-User Recommender System
0.00410
0.00405
0.00400
0.00395
0.00390
0.00385
0.00380
Mean Absolute Error (MAE)
10
15
20
25
30
35
40
45
50
N (Number of Similar Users)
Mean Absolute Error (MAE)



```
for K=10
MAE = 1.32421
for K=20
MAE = 0.89809
for K=30
MAE = 0.634112
for K=40
MAE = 1.36002
for K=50
MAE = 1.39866
```



11.The top 10 products identified based on the sum of ratings indicate the most popular or highly-rated products among users.



```
Top 10 products by User Sum Ratings:
asin
B00029MTMQ    10
B00008Z1QI    10
B00016556C    10
B00013BLEK    10
B00012F8AY    10
B0000UV2AW    10
B0000E3DQ7    10
B0000DZDHB    10
B0000ACCJA    10
B00009V2OV    10
Name: overall, dtype: int64
```