IR Assignment 3

Name: Drishti Singh
Roll No: MT23117

1. The 5-core electronics dataset was downloaded from the Small subset for experimentation, along with the meta-data.
2. All the rows corresponding to the “HeadPhones” product were taken out, using the asin values from the meta-data

- **Data Pre-processing**
  Missing values are handled by adding empty strings and empty lists and dictionaries in the 5-core headphone electronics dataset.
  Duplicates rows are removed from both the 5-core headphone electronics and metadata to clean the whole dataset.

- **Descriptive Statistics are calculated accordingly on the dataset**
  Descriptive Statistics of the product as : -
  a. Number of Reviews.

```
Number of Reviews are 625335
```

  b. Average Rating Score.

```
Average Rating is 4.11
```

  c. Number of Unique Products

```
Number of Unique Products: 4497
```

  d. Number of Good Rating (Ratings >= 3 are good)

```
Number of Good Ratings (>= 3) is 535699
```

  e. Number of Bad Ratings

```
Number of Bad Ratings (< 3) is 89732
```

  f. Number of Reviews corresponding to each Rating

```
Number of reviews per rating is as follows
 overall
1.0     49079
2.0     40653
3.0     60768
4.0    117875
5.0    357056
```

- **Text Preprocessing:**
  Text is pre-processed as follows:
  a. HTML Tags removed using Beautiful Soup library
  b. Accented characters are removed
  c. Acronyms (generated potential from chatgpt) are expanded
  d. Special characters are removed using appropriate regex
  e. Lemmatization done to bring words in their base form
  f. Text Normalization done by converting the whole text in lowercase and removing stop words.

- **Exploratory Data Analysis:**
  To get the brands for the products, the asin and brand of the products were taken from metadata and then the headphone dataset is joined with this to get the brand in the new dataset.
  Then the following things are analyzed;
  a. Top 20 most reviewed brands in the Headphone category

```
Top 20 most reviewed brands are brand
Sony              37457
Sennheiser        22977
Plantronics       11948
Bose              11583
Panasonic          8519
Skullcandy         7897
Mpow               7605
JLAB               7513
Roku               7469
JVC                7114
TaoTronics         7065
Samsung            7065
Audio-Technica     7056
Philips            7052
Koss               7002
Kinivo             6444
Apple              6375
Etre Jeune         5970
AmazonBasics       5898
LG                 5695
Name: reviewText, dtype: int64
```

b. Top 20 least reviewed brands in Headphone

```
Top 20 least reviewed brands are brand
Honda                    1
Digital Antenna          1
Zelco Industries, Inc    3
Fred & Friends           3
DSI                      3
NOIZY Brands             3
DetectorPro              4
SOUND-SQUARED CO.        4
SmartDisk                4
Replug                   4
YooZoo                   4
MEDca                    5
Abusun                   5
Basstyle                 5
BearsFire                5
ATEX                     5
ABC(TM)                  5
lexastech                5
meda one                 5
EveryMarket              5
```

c. Most positively reviewed Headphone

```
Highest rated brand: 4 in 1 Charger
Average rating: 5.0
```

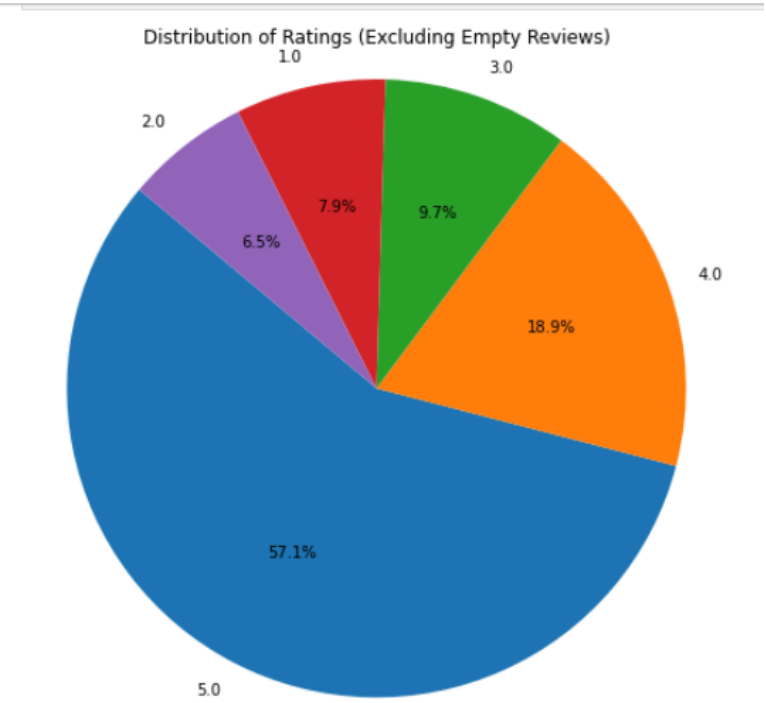
d. Count of ratings for the product over 5 consecutive years

```
Count of ratings for the last 5 years:
reviewTime
2014     93590
2015    145260
2016    143320
2017     80955
2018     32366
```

e. Word Cloud for ‘Good’ and ‘Bad’ ratings. (Good ratings are reviews with rating >=3)


```
Most commonly used words for good reviews:
sound - 355476
use - 278604
headphones - 263165
great - 233230
good - 222220
work - 188505
get - 174081
quality - 165244
like - 162151
one - 133130
------------------------------------------------
Most commonly used words for bad reviews:
sound - 55101
work - 43919
use - 41971
headphones - 37902
get - 37185
one - 27859
quality - 26193
buy - 25795
would - 25068
good - 24909
```

f. Pie chart for Distribution of Ratings vs. the No. of Reviews



g. Year in which Headphone got maximum reviews

```
Year with maximum reviews: 2015
```

h. Year which has the highest number of Customers

```
Year with the highest number of customers: 2016
```

- From the dataset, BERT model by Hugging Face is used to get the embeddings of the reviews and saved in separate numpy files. I had used the Word2Vec model, but due to the enormous size of the dataset, it failed to create embeddings in sufficient time.

- **Machine Learning Models:**
  The 5-core headphone dataset is divided into training and test set in 75:25 ratio. Following models are used to train the embeddings:

  a. Logistic Regression

```
Model: Logistic Regression
Class: Good
  Precision: 0.8309923594271857
  Recall: 0.38441558441558443
  F1-score: 0.6256752462662853

Class: Average
  Precision: 0.9577620755114955
  Recall: 0.05597579425113464
  F1-score: 0.5207617032531077

Class: Bad
  Precision: 0.8898851081551162
  Recall: 0.09772202046880159
  F1-score: 0.568418013856813
```

b. Decision Tree

```
Model: Decision Tree
Class: Good
  Precision: 0.8028267411865864
  Recall: 0.15125173852573018
  F1-score: 0.3142125480153649

Class: Average
  Precision: 0.7877557477325459
  Recall: 0.1645234493192133
  F1-score: 0.32451732345940226

Class: Bad
  Precision: 0.7952198445651018
  Recall: 0.1576086956521739
  F1-score: 0.3192818110850898
```

c. Random Forest

```
Model: Random Forest
Class: Good
  Precision: 0.7703503822723742
  Recall: 0.6774193548387096
  F1-score: 0.7638248847926268

Class: Average
  Precision: 0.9935667580679182
  Recall: 0.01588502269288956
  F1-score: 0.17535043639248876

Class: Bad
  Precision: 0.8678350183082697
  Recall: 0.03104212860310421
  F1-score: 0.2852226285222629
```

d. KNN

```
Model: KNN
Class: Good
  Precision: 0.8009916302765647
  Recall: 0.18826619964973731
  F1-score: 0.5196641626159965

Class: Average
  Precision: 0.9285488293608943
  Recall: 0.08131618759455371
  F1-score: 0.31102882835228773

Class: Bad
  Precision: 0.8600664257106574
  Recall: 0.11357633386159537
  F1-score: 0.38914626075446723
```

e. XGBoost

```
Model: XGBoost
Class: Good
  Precision: 0.30327868852459017
  Recall: 0.6301174795300819
  F1-score: 0.8201134419735302

Class: Average
  Precision: 0.041981845688350984
  Recall: 0.46813012430573925
  F1-score: 0.960662307530057

Class: Bad
  Precision: 0.07375415282392027
  Recall: 0.5371775417298938
  F1-score: 0.884841420175822
```

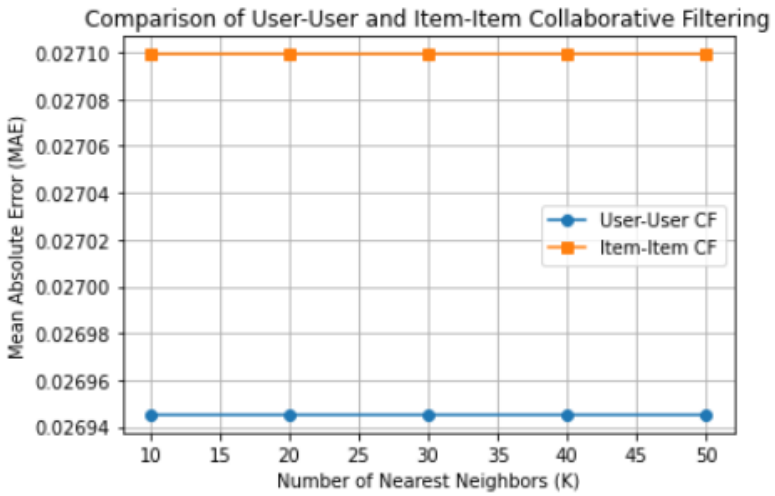
- **Collaborative Filtering:**
  a. The data is created by creating a user - item matrix, user being the reviewerID, item being the asin and ratings being the overall column.
  b. Find the top N similar users, by using cosine similarity. N = 10, 20, 30, 40, 50

  Displaying for only 1 user:

```
ReviewerID A1010IB6ZI4YWK:
  Top 10 similar users: ['AY88L1GAS7HN7', 'ATUL99V1C0U6X', 'AS27K6R1MGMDY', 'ARJDGQ1C1Z1FT', 'A
MBX0TR9DL3JS', 'AID92BWEGPK99', 'AFTD9I043S1KP', 'AF8R1TOUWM3VO', 'AAZRAJZY9BLZV', 'A3RVZDHWG8C
GEJ']
--------------------------------------------------------------------------------------------
-------
  Top 20 similar users: ['AY88L1GAS7HN7', 'ATUL99V1C0U6X', 'AS27K6R1MGMDY', 'ARJDGQ1C1Z1FT', 'A
MBX0TR9DL3JS', 'AID92BWEGPK99', 'AFTD9I043S1KP', 'AF8R1TOUWM3VO', 'AAZRAJZY9BLZV', 'A3RVZDHWG8C
GEJ', 'A3PCG3K2WYHFUE', 'A3NI5J2G7BNVS9', 'A3MCV7ORKJ7GQ2', 'A3IARFW5PBDOCT', 'A3HXPOYTN8B4SU',
'A3FCON8N6PJIZB', 'A3ACFC6DQQLIQT', 'A39VIFSCU077S1', 'A38RQFVQ1AKJQQ', 'A36BC0YFDBNB5X']
--------------------------------------------------------------------------------------------
-------
  Top 30 similar users: ['AY88L1GAS7HN7', 'ATUL99V1C0U6X', 'AS27K6R1MGMDY', 'ARJDGQ1C1Z1FT', 'A
MBX0TR9DL3JS', 'AID92BWEGPK99', 'AFTD9I043S1KP', 'AF8R1TOUWM3VO', 'AAZRAJZY9BLZV', 'A3RVZDHWG8C
GEJ', 'A3PCG3K2WYHFUE', 'A3NI5J2G7BNVS9', 'A3MCV7ORKJ7GQ2', 'A3IARFW5PBDOCT', 'A3HXPOYTN8B4SU',
'A3FCON8N6PJIZB', 'A3ACFC6DQQLIQT', 'A39VIFSCU077S1', 'A38RQFVQ1AKJQQ', 'A36BC0YFDBNB5X', 'A34J
4E1N58BZOQ', 'A2Z6U9LVEO5BI9', 'A2YBAEFSOV1NC5', 'A2Y5C96KFQS9KT', 'A2VF9N3SBYTMXR', 'A2RKGF62S
YF6KT', 'A2RGD32TFI03S3', 'A2NO2IF75CI811', 'A2MTNRYCL8ZQ3J', 'A2HLEFPRRZXV8F']
--------------------------------------------------------------------------------------------
-------
```

```
  Top 40 similar users: ['AY88L1GAS7HN7', 'ATUL99V1C0U6X', 'AS27K6R1MGMDY', 'ARJDGQ1C1Z1FT', 'A
MBX0TR9DL3JS', 'AID92BWEGPK99', 'AFTD9I043S1KP', 'AF8R1TOUWM3VO', 'AAZRAJZY9BLZV', 'A3RVZDHWG8C
GEJ', 'A3PCG3K2WYHFUE', 'A3NI5J2G7BNVS9', 'A3MCV7ORKJ7GQ2', 'A3IARFW5PBDOCT', 'A3HXPOYTN8B4SU',
'A3FCON8N6PJIZB', 'A3ACFC6DQQLIQT', 'A39VIFSCU077S1', 'A38RQFVQ1AKJQQ', 'A36BC0YFDBNB5X', 'A34J
4E1N58BZOQ', 'A2Z6U9LVEO5BI9', 'A2YBAEFSOV1NC5', 'A2Y5C96KFQS9KT', 'A2VF9N3SBYTMXR', 'A2RKGF62S
YF6KT', 'A2RGD32TFI03S3', 'A2NO2IF75CI811', 'A2MTNRYCL8ZQ3J', 'A2HLEFPRRZXV8F', 'A2GW2KWH88WX1'
, 'A2C159Y83QNNIE', 'A299MRB9O6GWDE', 'A28WDN37ZUV1IA', 'A212PQ0HQPNNWM', 'A1WVXHK1QH1DU2', 'A1
V3B17DOW1XG8', 'A1R299WQB5OYZI', 'A1OW1ESILQ5BS7', 'A1L5U2FVCPSVOB']
--------------------------------------------------------------------------------------------
-------
  Top 50 similar users: ['AY88L1GAS7HN7', 'ATUL99V1C0U6X', 'AS27K6R1MGMDY', 'ARJDGQ1C1Z1FT', 'A
MBX0TR9DL3JS', 'AID92BWEGPK99', 'AFTD9I043S1KP', 'AF8R1TOUWM3VO', 'AAZRAJZY9BLZV', 'A3RVZDHWG8C
GEJ', 'A3PCG3K2WYHFUE', 'A3NI5J2G7BNVS9', 'A3MCV7ORKJ7GQ2', 'A3IARFW5PBDOCT', 'A3HXPOYTN8B4SU',
'A3FCON8N6PJIZB', 'A3ACFC6DQQLIQT', 'A39VIFSCU077S1', 'A38RQFVQ1AKJQQ', 'A36BC0YFDBNB5X', 'A34J
4E1N58BZOQ', 'A2Z6U9LVEO5BI9', 'A2YBAEFSOV1NC5', 'A2Y5C96KFQS9KT', 'A2VF9N3SBYTMXR', 'A2RKGF62S
YF6KT', 'A2RGD32TFI03S3', 'A2NO2IF75CI811', 'A2MTNRYCL8ZQ3J', 'A2HLEFPRRZXV8F', 'A2GW2KWH88WX1'
, 'A2C159Y83QNNIE', 'A299MRB9O6GWDE', 'A28WDN37ZUV1IA', 'A212PQ0HQPNNWM', 'A1WVXHK1QH1DU2', 'A1
V3B17DOW1XG8', 'A1R299WQB5OYZI', 'A1OW1ESILQ5BS7', 'A1L5U2FVCPSVOB', 'A1HWH7F472DHE7', 'A1E51BV
5BPAADB', 'A1DVB8QV6WPUY3', 'A1BUL0B5PD0OGU', 'A1B8RJA188FL9U', 'A17K2WV1IB9BP3', 'A143RNRZVBYC
9K', 'A130H624EX3T4N', 'AZZYJH0XNZ896', 'AZXGVCWZ4QBR7']
```

c. The data is divided into training and validation sets to evaluate the performance of the recommendation system.
d. K-fold cross-validation is used to ensure robustness and avoid overfitting.
e. Missing values in the matrix indicate items not rated or interacted with by users.
f. For each pair of users, cosine similarity is calculated using their rating vectors.
g. The k-nearest neighbors for each user are determined based on their similarity scores. These nearest neighbors are users who have similar preferences to the target user.
h. Missing values in the user-item rating matrix are predicted using collaborative filtering. For each missing value, the system looks at the ratings of the nearest neighbors who have rated that item and predicts the missing value based on their ratings.
i. The predicted ratings are compared to the actual ratings in the validation set and MAE (Mean Average Error is calculated accordingly).
j. The above process is done for both User - User recommender system and the Item - Item recommender system.
k. Graph is plotted by the same - MAE of Item - Item recommender system being a little higher than the User - User recommender system.



I. MAE is calculated for k = 10, 20, 30, 40, 50

- **TOP 10 products by User Sum Ratings**
  From the user - item matrix, the sum of ratings given by users for each product is calculated and top 10 are displayed.

```
Top 10 products by User Sum Ratings:
Product ID: B00001P4XH, Sum of Ratings: 342.0
Product ID: B00000J1F3, Sum of Ratings: 56.0
Product ID: B00000JBHP, Sum of Ratings: 55.800000000000004
Product ID: B00000JCTO, Sum of Ratings: 48.2
Product ID: B000001OMI, Sum of Ratings: 32.6
Product ID: B00000J1EJ, Sum of Ratings: 29.0
Product ID: 4126895493, Sum of Ratings: 28.8
Product ID: B00000I9HE, Sum of Ratings: 26.0
Product ID: B00000I9HF, Sum of Ratings: 21.8
Product ID: B00000J1ES, Sum of Ratings: 18.6
```