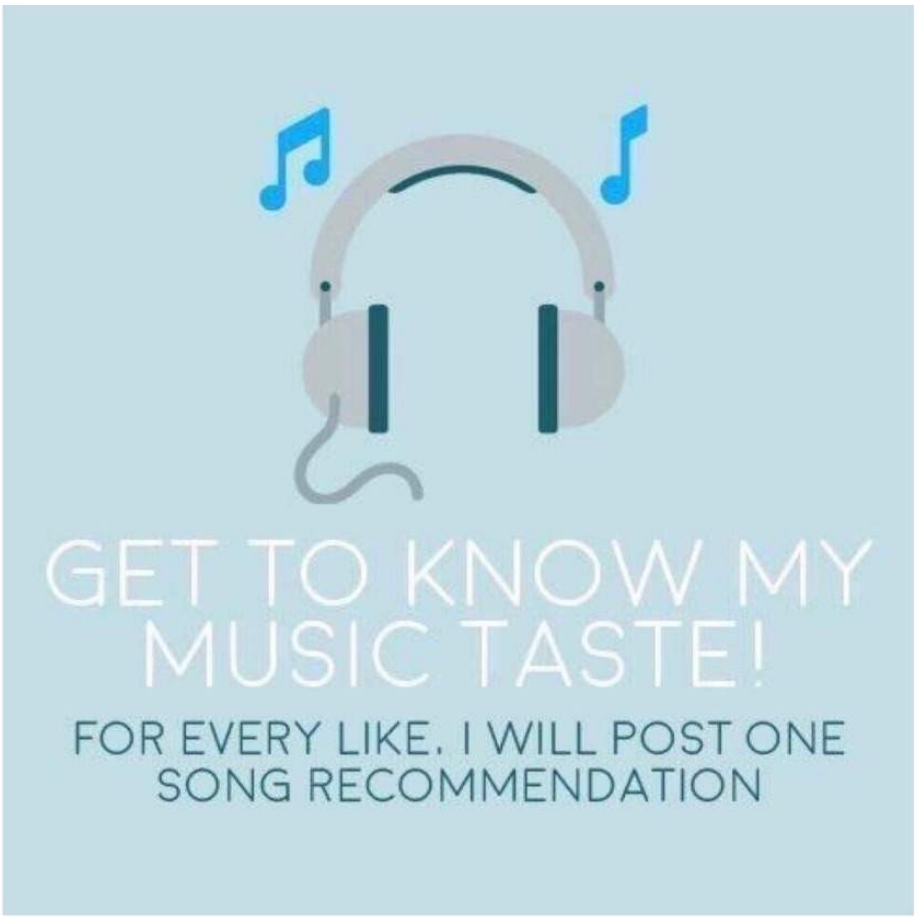
Song-Recommendation-ML



~~~Image has been taken from Google Image.

Recommendation of a song for the listener nased on gender, age, region, artist they like and many more.

Let's connect our jupyter notebook to jovian.

Problem Statement

I selected the 15th data set from the resources tab in Jovian. Link from where I downloaded the dataset: https://www.kaggle.com/c/MusicHackathon/data
~~~

This data has ratings given by the listeners, qualitative feedback, answers to the question on music and listeners demographics. We will use this dataset to get the rating of the test dataset.

It is a Regression type problem.

Installing the required libraries for making the model

```
!pip install plotly==5.11.0
```

```
Looking in indexes: https://pypi.org/simple, https://us-
python.pkg.dev/colab-wheels/public/simple/
Collecting plotly==5.11.0
  Downloading plotly-5.11.0-py2.py3-none-any.whl (15.3 MB)
ent already satisfied: tenacity>=6.2.0 in
/usr/local/lib/python3.8/dist-packages (from plotly==5.11.0) (8.1.0)
Installing collected packages: plotly
  Attempting uninstall: plotly
    Found existing installation: plotly 5.5.0
    Uninstalling plotly-5.5.0:
      Successfully uninstalled plotly-5.5.0
Successfully installed plotly-5.11.0
```

```
import matplotlib
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
%matplotlib inline
```

```
sns.set_style('darkgrid')
matplotlib.rcParams['font.size'] = 16
matplotlib.rcParams['figure.figsize'] = (14, 10)
matplotlib.rcParams['figure.facecolor'] = '#00000000'
```

```
!pip install opendatasets
```

```
Looking in indexes: https://pypi.org/simple, https://us-
python.pkg.dev/colab-wheels/public/simple/
Collecting opendatasets
  Downloading opendatasets-0.1.22-py3-none-any.whl (15 kB)
Requirement already satisfied: tqdm in /usr/local/lib/python3.8/dist-
packages (from opendatasets) (4.64.1)
Requirement already satisfied: kaggle in
/usr/local/lib/python3.8/dist-packages (from opendatasets) (1.5.12)
Requirement already satisfied: click in /usr/local/lib/python3.8/dist-
packages (from opendatasets) (7.1.2)
Requirement already satisfied: python-slugify in
/usr/local/lib/python3.8/dist-packages (from kaggle->opendatasets)
(7.0.0)
Requirement already satisfied: certifi in
/usr/local/lib/python3.8/dist-packages (from kaggle->opendatasets)
```

```
(2022.12.7)
Requirement already satisfied: six>=1.10 in
/usr/local/lib/python3.8/dist-packages (from kaggle->opendatasets)
(1.15.0)
Requirement already satisfied: requests in
/usr/local/lib/python3.8/dist-packages (from kaggle->opendatasets)
(2.23.0)
Requirement already satisfied: urllib3 in
/usr/local/lib/python3.8/dist-packages (from kaggle->opendatasets)
(1.24.3)
Requirement already satisfied: python-dateutil in
/usr/local/lib/python3.8/dist-packages (from kaggle->opendatasets)
(2.8.2)
Requirement already satisfied: text-unidecode>=1.3 in
/usr/local/lib/python3.8/dist-packages (from python-slugify->kaggle-
>opendatasets) (1.3)
Requirement already satisfied: chardet<4,>=3.0.2 in
/usr/local/lib/python3.8/dist-packages (from requests->kaggle-
>opendatasets) (3.0.4)
Requirement already satisfied: idna<3,>=2.5 in
/usr/local/lib/python3.8/dist-packages (from requests->kaggle-
>opendatasets) (2.10)
Installing collected packages: opendatasets
Successfully installed opendatasets-0.1.22
```

```
import os
import opendatasets as od
import pandas as pd
import numpy as np
pd.set_option("display.max_columns", 120)
pd.set_option("display.max_rows", 120)
```

Downloading data set from Kaggle in the notebook

```
od.download('https://www.kaggle.com/c/MusicHackathon/data')
```

```
Downloading MusicHackathon.zip to ./MusicHackathon

100%|██████████| 6.62M/6.62M [00:00<00:00, 47.9MB/s]


Extracting archive ./MusicHackathon/MusicHackathon.zip to
./MusicHackathon
```

```
os.listdir('MusicHackathon')
```

```
['UserKey.csv',
 'global_mean_benchmark.csv',
 'words.csv',
 'tracks_mean_benchmark.csv',
```

```
 'sample.r',
 'artists_mean_benchmark.csv',
 'users_mean_benchmark.csv',
 'test.csv',
 'logo_greenplum_main.png',
 'users.csv',
 'train.csv']
```

Converting the dataset to dataframe

```
train_df = pd.read_csv('./MusicHackathon/train.csv')
test_df = pd.read_csv('./MusicHackathon/test.csv')
words_df = pd.read_csv('./MusicHackathon/words.csv', encoding = "ISO-
8859-1")
users_df = pd.read_csv('./MusicHackathon/users.csv')
```

```
train_df
```

```
        Artist  Track   User  Rating  Time
0           40    179  47994       9    17
1            9     23   8575      58     7
2           46    168  45475      13    16
3           11    153  39508      42    15
4           14     32  11565      54    19
...        ...    ...    ...     ...   ...
188685       0      3   1278      29     6
188686       1      6   2839      30    18
188687      10    142  35756      61    12
188688      22     54  20163      46    21
188689      47    171  45580      12     4

[188690 rows x 5 columns]
```

```
test_df
```

```
        Artist  Track   User  Time
0            1      6   3475    18
1            6    149  39210    15
2           40    177  47861    17
3           31     79  27413    11
4           26     66  23232    22
...        ...    ...    ...   ...
125789      14     95  30004    23
125790      10     25   8186     7
125791      40    146  38180    13
125792      22    113  32918     0
125793       2     70  24231    22

[125794 rows x 4 columns]
```

```
words_df
```

```
        Artist   User                                HEARD_OF  \
0           47  45969                                Heard of
1           35  29118                           Never heard of
2           14  31544                                Heard of
3           23  18085                           Never heard of
4           23  18084                           Never heard of
...        ...    ...                                     ...
118296       4   3932      Heard of and listened to music EVER
118297       4   3935      Heard of and listened to music EVER
118298      12  11216  Heard of and listened to music RECENTLY
118299      33  35142      Heard of and listened to music EVER
118300       4   3915      Heard of and listened to music EVER

                   OWN_ARTIST_MUSIC  LIKE_ARTIST  Uninspired
Sophisticated  \
0                               NaN          NaN         NaN
0.0
1                               NaN          NaN         0.0
NaN
2                               NaN          NaN         0.0
NaN
3                               NaN          NaN         NaN
NaN
4                               NaN          NaN         NaN
NaN
...                             ...          ...         ...
...
118296  Own a little of their music         26.0         NaN
NaN
118297  Own a little of their music         30.0         NaN
NaN
118298      Own none of their music         71.0         NaN
NaN
118299      Own none of their music         31.0         NaN
NaN
118300  Own a little of their music         46.0         NaN
NaN

        Aggressive  Edgy  Sociable  Laid back  Wholesome  Uplifting  \
0              NaN     0       0.0        0.0        NaN        0.0
1              0.0     0       NaN        NaN        NaN        NaN
2              0.0     0       NaN        NaN        NaN        NaN
3              0.0     0       NaN        NaN        NaN        NaN
4              0.0     0       NaN        NaN        NaN        NaN
...            ...   ...       ...        ...        ...        ...
118296         0.0     0       NaN        NaN        NaN        NaN
118297         0.0     0       NaN        NaN        NaN        NaN
118298         0.0     0       NaN        NaN        NaN        NaN
118299         0.0     0       NaN        NaN        NaN        NaN
118300         0.0     0       NaN        NaN        NaN        NaN
```

```
        Intriguing  Legendary  Free  Thoughtful  Outspoken  Serious  \
0              0.0        NaN   0.0           0        0.0      NaN
1              NaN        NaN   NaN           0        NaN      0.0
2              NaN        NaN   NaN           1        NaN      0.0
3              NaN        NaN   NaN           0        NaN      0.0
4              NaN        NaN   NaN           0        NaN      0.0
...            ...        ...   ...         ...        ...      ...
118296         NaN        NaN   NaN           0        NaN      0.0
118297         NaN        NaN   NaN           0        NaN      0.0
118298         NaN        NaN   NaN           0        NaN      0.0
118299         NaN        NaN   NaN           0        NaN      0.0
118300         NaN        NaN   NaN           0        NaN      0.0

        Good lyrics  Unattractive  Confident  Old  Youthful  Boring
Current  \
0               NaN           NaN        NaN  NaN       0.0     1.0
0
1               0.0           0.0        0.0  NaN       0.0     0.0
0
2               0.0           0.0        0.0  NaN       0.0     0.0
0
3               0.0           0.0        0.0  NaN       0.0     1.0
0
4               0.0           0.0        0.0  NaN       0.0     0.0
0
...             ...           ...        ...  ...       ...     ...
...
118296          0.0           0.0        0.0  NaN       0.0     NaN
0
118297          0.0           0.0        0.0  NaN       0.0     NaN
0
118298          0.0           0.0        0.0  NaN       0.0     0.0
1
118299          0.0           0.0        0.0  NaN       0.0     0.0
1
118300          0.0           0.0        0.0  NaN       0.0     NaN
0

        Colourful  Stylish  Cheap  Irrelevant  Heartfelt  Calm
Pioneer  \
0             0.0        0    NaN         NaN        0.0   NaN
NaN
1             NaN        0    0.0         0.0        NaN   0.0
NaN
2             NaN        0    0.0         0.0        NaN   1.0
NaN
3             NaN        0    0.0         NaN        NaN   0.0
NaN
4             NaN        0    0.0         NaN        NaN   0.0
```

```
NaN
...           ...      ...    ...         ...        ...   ...      ..
.
118296        NaN        0    0.0         NaN        NaN   0.0
NaN
118297        NaN        0    1.0         NaN        NaN   0.0
NaN
118298        NaN        0    0.0         NaN        NaN   0.0
NaN
118299        NaN        0    0.0         NaN        NaN   0.0
NaN
118300        NaN        0    0.0         NaN        NaN   0.0
NaN

        Outgoing  Inspiring  Beautiful  Fun  Authentic  Credible  Way
out  \
0            NaN        NaN          0    0          0         0
0.0
1            0.0        0.0          0    1          0         0
NaN
2            0.0        0.0          1    0          0         0
NaN
3            0.0        0.0          0    0          0         0
NaN
4            0.0        0.0          0    0          0         0
NaN
...          ...        ...        ...  ...        ...       ...
...
118296       0.0        0.0          0    0          0         0
NaN
118297       0.0        0.0          0    0          0         0
NaN
118298       0.0        0.0          1    1          0         0
NaN
118299       0.0        0.0          0    0          0         0
NaN
118300       0.0        0.0          0    0          0         0
NaN

        Cool  Catchy  Sensitive  Mainstream  Superficial  Annoying
Dark  \
0          0     0.0        NaN         NaN          NaN       NaN
NaN
1          0     1.0        0.0         0.0          0.0       0.0
NaN
2          0     0.0        0.0         0.0          0.0       0.0
NaN
3          0     0.0        0.0         0.0          0.0       NaN
NaN
4          0     0.0        0.0         0.0          0.0       NaN
```

```
NaN
...      ...     ...        ...         ...          ...       ...   .
..
118296     0     0.0        0.0         NaN          0.0       NaN
NaN
118297     0     0.0        0.0         NaN          0.0       NaN
NaN
118298     0     1.0        0.0         0.0          0.0       NaN
NaN
118299     1     0.0        0.0         NaN          0.0       NaN
NaN
118300     0     0.0        0.0         NaN          0.0       NaN
NaN

        Passionate  Not authentic  Good Lyrics  Background
Timeless  \
0                0            NaN          0.0         0.0         0

1                0            0.0          NaN         NaN         0

2                0            0.0          NaN         NaN         0

3                0            NaN          NaN         NaN         0

4                0            NaN          NaN         NaN         0

...            ...            ...          ...         ...       ...

118296           0            NaN          NaN         NaN         0

118297           0            NaN          NaN         NaN         0

118298           0            NaN          NaN         NaN         0

118299           0            NaN          NaN         NaN         0

118300           0            NaN          NaN         NaN         0


        Depressing  Original  Talented  Worldly  Distinctive
Approachable  \
0              NaN         0         0      NaN            0
0
1              0.0         0         0      NaN            1
0
2              0.0         0         1      NaN            0
0
3              0.0         0         0      NaN            0
0
```

```
4              0.0         0         0      NaN            0
0
...            ...       ...       ...      ...          ...
...
118296         0.0         0         0      NaN            0
0
118297         0.0         1         0      NaN            0
0
118298         0.0         0         0      NaN            0
1
118299         0.0         0         0      NaN            0
0
118300         0.0         1         0      NaN            1
0

        Genius  Trendsetter  Noisy  Upbeat  Relatable  Energetic
Exciting  \
0          0.0            0    NaN     0.0        NaN          0
0.0
1          NaN            0    0.0     0.0        0.0          0
NaN
2          NaN            0    0.0     0.0        0.0          0
NaN
3          NaN            0    0.0     0.0        0.0          0
NaN
4          NaN            0    0.0     0.0        0.0          0
NaN
...        ...          ...    ...     ...        ...        ...
...
118296     NaN            0    0.0     0.0        NaN          0
NaN
118297     NaN            0    0.0     1.0        NaN          0
NaN
118298     NaN            0    0.0     0.0        0.0          0
NaN
118299     NaN            0    0.0     0.0        NaN          1
NaN
118300     NaN            0    0.0     0.0        NaN          1
NaN

        Emotional  Nostalgic  None of these  Progressive  Sexy
Over  \
0             0.0        NaN              0          NaN     0   NaN

1             NaN        NaN              0          NaN     0   0.0

2             NaN        NaN              0          NaN     0   0.0

3             NaN        NaN              0          NaN     0   0.0
```

```
4             NaN        NaN              1          NaN     0   0.0

...           ...        ...            ...          ...   ...   ...

118296        NaN        NaN              0          NaN     0   0.0

118297        NaN        NaN              0          NaN     0   0.0

118298        NaN        NaN              0          NaN     0   0.0

118299        NaN        NaN              0          NaN     0   0.0

118300        NaN        NaN              0          NaN     0   0.0


        Rebellious  Fake  Cheesy  Popular  Superstar  Relaxed
Intrusive  \
0              0.0   NaN     NaN      0.0        NaN      0.0
NaN
1              NaN   0.0     0.0      NaN        0.0      NaN
0.0
2              NaN   0.0     0.0      NaN        0.0      NaN
0.0
3              NaN   0.0     0.0      NaN        0.0      NaN
NaN
4              NaN   0.0     0.0      NaN        0.0      NaN
NaN
...            ...   ...     ...      ...        ...      ...        .
..
118296         NaN   0.0     0.0      NaN        NaN      NaN
NaN
118297         NaN   0.0     0.0      NaN        NaN      NaN
NaN
118298         NaN   0.0     0.0      NaN        0.0      NaN
NaN
118299         NaN   0.0     0.0      NaN        NaN      NaN
NaN
118300         NaN   0.0     0.0      NaN        NaN      NaN
NaN

        Unoriginal  Dated  Iconic  Unapproachable  Classic  Playful
Arrogant  \
0              NaN    0.0     NaN             NaN      0.0      NaN
NaN
1              0.0    0.0     NaN             0.0      0.0      0.0
0.0
2              0.0    0.0     NaN             0.0      0.0      0.0
0.0
3              0.0    0.0     NaN             0.0      0.0      0.0
```

```
0.0
4              0.0    0.0     NaN             0.0      0.0      0.0
0.0
...            ...    ...     ...             ...      ...      ...
...
118296         0.0    1.0     NaN             0.0      0.0      0.0
0.0
118297         0.0    0.0     NaN             0.0      0.0      0.0
0.0
118298         0.0    0.0     NaN             0.0      0.0      1.0
0.0
118299         0.0    0.0     NaN             0.0      NaN      0.0
1.0
118300         0.0    0.0     NaN             0.0      0.0      1.0
0.0

        Warm  Soulful  Unnamed: 87
0          0      0.0          NaN
1          0      NaN          NaN
2          0      NaN          NaN
3          0      NaN          NaN
4          0      NaN          NaN
...      ...      ...          ...
118296     0      NaN          NaN
118297     0      NaN          NaN
118298     0      NaN          NaN
118299     0      NaN          NaN
118300     0      NaN          NaN

[118301 rows x 88 columns]

users_df

       RESPID  GENDER   AGE                                WORKING
REGION  \
0       36927  Female  60.0                                  Other
South
1        3566  Female  36.0  Full-time housewife / househusband
South
2       20054  Female  52.0           Employed 30+ hours a week
Midlands
3       41749  Female  40.0        Employed 8-29 hours per week
South
4       23108  Female  16.0                    Full-time student
North
...       ...     ...   ...                                  ...
...
48640   19361    Male  48.0                       Self-employed
Midlands
48641   17639  Female  60.0  Full-time housewife / househusband
```

```
Midlands
48642   28753  Female  25.0           Employed 30+ hours a week
Midlands
48643   26197    Male  44.0           Employed 30+ hours a week
Midlands
48644   16225  Female  43.0                               NaN
North

                                                      MUSIC
LIST_OWN  \
0      Music is important to me but not necessarily m...              1
hour
1      Music is important to me but not necessarily m...              1
hour
2      I like music but it does not feature heavily i...              1
hour
3       Music means a lot to me and is a passion of mine             2
hours
4       Music means a lot to me and is a passion of mine             3
hours
...                                                  ...
...
48640  I like music but it does not feature heavily i...  Less than an
hour
48641   Music means a lot to me and is a passion of mine             2
hours
48642   Music means a lot to me and is a passion of mine             2
hours
48643   Music means a lot to me and is a passion of mine             2
hours
48644  I like music but it does not feature heavily i...
NaN

               LIST_BACK    Q1    Q2    Q3    Q4    Q5    Q6     Q7
Q8  \
0                    NaN  49.0  50.0  49.0  50.0  32.0  33.0   32.0
0.0
1                 1 hour  55.0  55.0  62.0   9.0   9.0   9.0   10.0
11.0
2      Less than an hour  11.0  50.0   9.0   8.0  45.0  10.0   30.0
29.0
3                3 hours  81.0  80.0  88.0  88.0  31.0  31.0   51.0
30.0
4                6 hours  76.0  79.0  78.0  73.0  71.0  68.0   73.0
67.0
...                  ...   ...   ...   ...   ...   ...   ...    ...
...
48640            2 hours   9.0  73.0  33.0   6.0  10.0  68.0   51.0
52.0
48641             1 hour  26.0  50.0  49.0  58.0  59.0  48.0    6.0
```

```
5.0
48642            6 hours  89.0  89.0  89.0   6.0   6.0  51.0   26.0
5.0
48643            4 hours  95.0  97.0  97.0  98.0  97.0  99.0  100.0
99.0
48644                  2  49.0  48.0  50.0  51.0  49.0  25.0   45.0
51.0

          Q9   Q10   Q11   Q12   Q13   Q14   Q15   Q16    Q17   Q18
Q19
0       74.0  50.0  50.0  71.0  52.0  71.0   9.0   7.0   72.0  49.0
26.0
1       55.0  12.0  65.0  65.0  80.0  79.0  51.0  31.0   68.0  54.0
33.0
2        8.0  50.0  94.0  51.0  74.0  66.0  27.0  46.0   73.0   8.0
31.0
3        8.0  76.0  74.0  64.0  73.0  85.0  61.0  77.0   76.0  78.0
88.0
4       31.0  56.0  13.0  82.0  79.0  68.0  71.0   NaN   86.0  80.0
32.0
...      ...   ...   ...   ...   ...   ...   ...   ...    ...   ...
...
48640   93.0  53.0  74.0  36.0  13.0  38.0  12.0  10.0   50.0  10.0
28.0
48641   88.0  58.0  62.0  79.0  17.0  24.0  30.0   6.0   73.0  20.0
21.0
48642    0.0  70.0  70.0  70.0  51.0  70.0  70.0   NaN  100.0  70.0
69.0
48643  100.0  97.0  98.0  99.0  97.0  99.0  99.0  99.0  100.0  91.0
96.0
48644   40.0  10.0  69.0  70.0  53.0  54.0  10.0   4.0    7.0   NaN
NaN

[48645 rows x 27 columns]

words_df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 118301 entries, 0 to 118300
Data columns (total 88 columns):
 #   Column            Non-Null Count   Dtype
---  ------            --------------   -----
 0   Artist            118301 non-null  int64
 1   User              118301 non-null  int64
 2   HEARD_OF          118277 non-null  object
 3   OWN_ARTIST_MUSIC  33507 non-null   object
 4   LIKE_ARTIST       33308 non-null   float64
 5   Uninspired        26154 non-null   float64
 6   Sophisticated     20724 non-null   float64
 7   Aggressive        97577 non-null   float64
```

8	Edgy	118301 non-null	int64
9	Sociable	20724 non-null	float64
10	Laid back	20724 non-null	float64
11	Wholesome	1040 non-null	float64
12	Uplifting	20724 non-null	float64
13	Intriguing	20724 non-null	float64
14	Legendary	1040 non-null	float64
15	Free	20724 non-null	float64
16	Thoughtful	118301 non-null	int64
17	Outspoken	20724 non-null	float64
18	Serious	97577 non-null	float64
19	Good lyrics	97577 non-null	float64
20	Unattractive	97577 non-null	float64
21	Confident	97577 non-null	float64
22	Old	1040 non-null	float64
23	Youthful	117261 non-null	float64
24	Boring	87080 non-null	float64
25	Current	118301 non-null	int64
26	Colourful	20724 non-null	float64
27	Stylish	118301 non-null	int64
28	Cheap	97577 non-null	float64
29	Irrelevant	26154 non-null	float64
30	Heartfelt	20724 non-null	float64
31	Calm	97577 non-null	float64
32	Pioneer	1040 non-null	float64
33	Outgoing	97577 non-null	float64
34	Inspiring	97577 non-null	float64
35	Beautiful	118301 non-null	int64
36	Fun	118301 non-null	int64
37	Authentic	118301 non-null	int64
38	Credible	118301 non-null	int64
39	Way out	20724 non-null	float64
40	Cool	118301 non-null	int64
41	Catchy	117261 non-null	float64
42	Sensitive	97577 non-null	float64
43	Mainstream	46254 non-null	float64
44	Superficial	97577 non-null	float64
45	Annoying	26154 non-null	float64
46	Dark	1040 non-null	float64
47	Passionate	118301 non-null	int64
48	Not authentic	26154 non-null	float64
49	Good Lyrics	20724 non-null	float64
50	Background	20724 non-null	float64
51	Timeless	118301 non-null	int64
52	Depressing	97577 non-null	float64
53	Original	118301 non-null	int64
54	Talented	118301 non-null	int64
55	Worldly	1040 non-null	float64
56	Distinctive	118301 non-null	int64
57	Approachable	118301 non-null	int64

```
 58  Genius            20724 non-null   float64
 59  Trendsetter       118301 non-null  int64
 60  Noisy             97577 non-null   float64
 61  Upbeat            117261 non-null  float64
 62  Relatable         46254 non-null   float64
 63  Energetic         118301 non-null  int64
 64  Exciting          20724 non-null   float64
 65  Emotional         20724 non-null   float64
 66  Nostalgic         1040 non-null    float64
 67  None of these     118301 non-null  int64
 68  Progressive       1040 non-null    float64
 69  Sexy              118301 non-null  int64
 70  Over              90157 non-null   float64
 71  Rebellious        20724 non-null   float64
 72  Fake              97577 non-null   float64
 73  Cheesy            97577 non-null   float64
 74  Popular           19684 non-null   float64
 75  Superstar         46254 non-null   float64
 76  Relaxed           20724 non-null   float64
 77  Intrusive         26154 non-null   float64
 78  Unoriginal        97577 non-null   float64
 79  Dated             117261 non-null  float64
 80  Iconic            1040 non-null    float64
 81  Unapproachable    97577 non-null   float64
 82  Classic           105235 non-null  float64
 83  Playful           97577 non-null   float64
 84  Arrogant          97577 non-null   float64
 85  Warm              118301 non-null  int64
 86  Soulful           19684 non-null   float64
 87  Unnamed: 87       0 non-null       float64
dtypes: float64(64), int64(22), object(2)
memory usage: 79.4+ MB
```

Score to words DF

Now i will be giving score to 'words_df' by preprocessing the df.

The score system works like this:

- For each value 1 in the positive columns, we **add 1 point to the total score**
- For each value 1 in the negative columns, we **subtract 1 point to the total score**
- Any 0 and NaN value we **ignore as they are neutral**

```
positive_score = ['Sophisticated', 'Sociable', 'Laid back',
'Wholesome', 'Uplifting', 'Intriguing', 'Legendary', 'Free',
'Outspoken', 'Good lyrics', 'Confident', 'Youthful', 'Current',
'Colourful', 'Stylish', 'Heartfelt', 'Pioneer', 'Outgoing',
'Inspiring', 'Beautiful', 'Fun', 'Authentic', 'Credible', 'Way out',
'Cool', 'Catchy', 'Sensitive', 'Passionate', 'Good Lyrics',
'Timeless', 'Original', 'Talented', 'Distinctive', 'Approachable',
```

```
'Genius', 'Trendsetter', 'Upbeat', 'Relatable', 'Energetic',
'Exciting', 'Emotional', 'Nostalgic', 'Progressive', 'Sexy', 'Over',
'Popular', 'Superstar', 'Relaxed', 'Iconic', 'Classic', 'Playful',
'Warm', 'Soulful']

negative_score = ['Uninspired', 'Unattractive', 'Boring', 'Cheap',
'Irrelevant', 'Superficial', 'Annoying', 'Not authentic',
'Depressing', 'Noisy', 'Fake', 'Cheesy', 'Intrusive', 'Unoriginal',
'Dated', 'Unapproachable']

words_df['plus_score'] = words_df[positive_score].sum(axis=1)
words_df['minus_score'] = words_df[negative_score].sum(axis=1)
words_df['words_score'] = words_df['plus_score'] -
words_df['minus_score']

words_df[words_df.LIKE_ARTIST > 90].sample(15)
```

	Artist	User	HEARD_OF	OWN_ARTIST_MUSIC	LIKE_ARTIST	Uninspired
109843	4	38089	Heard of and listened to music RECENTLY	Own all or most of their music	100.0	NaN
28247	41	42540	Heard of and listened to music RECENTLY	Own all or most of their music	95.0	NaN
104792	4	36390	Heard of and listened to music RECENTLY	Own all or most of their music	91.0	NaN
57761	17	14331	Heard of and listened to music RECENTLY	Own a lot of their music	91.0	NaN
20595	32	25628	Heard of and listened to music RECENTLY	Own a lot of their music	100.0	NaN
62408	40	36255	Heard of and listened to music RECENTLY	Own all or most of their music	93.0	NaN
6553	43	42931	Heard of and listened to music RECENTLY	Own all or most of their music	94.0	NaN
105115	4	36778	Heard of and listened to music RECENTLY	Own all or most of their music	93.0	NaN
107250	22	32235	Listened to recently	Own all or most of their music	91.0	NaN
31728	10	10421	Heard of and listened to music RECENTLY	Own a lot of their music	93.0	NaN
72534	34	28566	Heard of and listened to music EVER	Own none of their music	100.0	0.0
30895	12	14055	Heard of and listened to music RECENTLY	Own a little of their music	100.0	NaN
75007	36	30841	Heard of and listened to music EVER	Own a lot of their music	91.0	0.0
61166	24	21099	Heard of and listened to music RECENTLY	Own all or most of their music	100.0	NaN
24462	4	2198	Heard of and listened to music RECENTLY	Own all or most of their music	91.0	NaN

	Sophisticated	Aggressive	Edgy	Sociable	Laid back	Wholesome
109843	NaN	0.0	1	NaN	NaN	NaN
28247	NaN	0.0	0	NaN	NaN	NaN
104792	NaN	0.0	0	NaN	NaN	NaN
57761	0.0	NaN	0	0.0	0.0	NaN
20595	NaN	0.0	1	NaN	NaN	NaN
62408	NaN	0.0	0	NaN	NaN	NaN
6553	NaN	0.0	1	NaN	NaN	NaN
105115	NaN	0.0	0	NaN	NaN	NaN
107250	0.0	NaN	0	0.0	0.0	0.0
31728	0.0	NaN	0	0.0	0.0	NaN
72534	NaN	0.0	0	NaN	NaN	NaN
30895	1.0	NaN	0	0.0	0.0	NaN
75007	NaN	0.0	0	NaN	NaN	NaN
61166	NaN	1.0	0	NaN	NaN	NaN
24462	NaN	0.0	0	NaN	NaN	NaN

	Uplifting	Intriguing	Legendary	Free	Thoughtful	Outspoken
109843	NaN	NaN	NaN	NaN	1	NaN
28247	NaN	NaN	NaN	NaN	0	NaN
104792	NaN	NaN	NaN	NaN	0	NaN
57761	1.0	0.0	NaN	1.0	0	0.0
20595	NaN	NaN	NaN	NaN	0	NaN
62408	NaN	NaN	NaN	NaN	0	NaN
6553	NaN	NaN	NaN	NaN	1	NaN

105115	NaN	NaN	NaN	NaN	1	NaN
107250	0.0	0.0	0.0	0.0	1	1.0
31728	0.0	0.0	NaN	0.0	0	0.0
72534	NaN	NaN	NaN	NaN	1	NaN
30895	0.0	0.0	NaN	1.0	1	0.0
75007	NaN	NaN	NaN	NaN	0	NaN
61166	NaN	NaN	NaN	NaN	0	NaN
24462	NaN	NaN	NaN	NaN	0	NaN

	Serious	Good lyrics	Unattractive	Confident	Old	Youthful	Boring	\
109843	0.0	1.0	0.0	0.0	NaN	0.0	0.0	
28247	0.0	1.0	0.0	0.0	NaN	0.0	NaN	
104792	0.0	0.0	0.0	0.0	NaN	0.0	0.0	
57761	NaN	NaN	NaN	NaN	NaN	1.0	0.0	
20595	0.0	1.0	0.0	1.0	NaN	0.0	0.0	
62408	0.0	0.0	0.0	0.0	NaN	0.0	0.0	
6553	1.0	1.0	0.0	1.0	NaN	0.0	NaN	
105115	0.0	1.0	0.0	1.0	NaN	0.0	0.0	
107250	NaN	NaN	NaN	NaN	0.0	NaN	NaN	
31728	NaN	NaN	NaN	NaN	NaN	0.0	0.0	
72534	0.0	0.0	0.0	0.0	NaN	0.0	0.0	
30895	NaN	NaN	NaN	NaN	NaN	1.0	0.0	
75007	0.0	0.0	0.0	0.0	NaN	0.0	0.0	
61166	0.0	0.0	0.0	0.0	NaN	0.0	0.0	
24462	0.0	1.0	0.0	1.0	NaN	0.0	NaN	

```
        Current  Colourful  Stylish  Cheap  Irrelevant  Heartfelt
Calm  \
109843        0        NaN        0    0.0         NaN        NaN
0.0
28247         0        NaN        0    0.0         NaN        NaN
0.0
104792        0        NaN        0    0.0         NaN        NaN
0.0
57761         1        1.0        0    NaN         NaN        0.0
NaN
20595         1        NaN        1    0.0         NaN        NaN
0.0
62408         1        NaN        0    0.0         NaN        NaN
0.0
6553          1        NaN        0    0.0         NaN        NaN
1.0
105115        0        NaN        0    0.0         NaN        NaN
0.0
107250        0        0.0        1    NaN         NaN        0.0
NaN
31728         0        0.0        0    NaN         NaN        0.0
NaN
72534         0        NaN        0    0.0         0.0        NaN
0.0
30895         1        1.0        1    NaN         NaN        1.0
NaN
75007         0        NaN        0    0.0         0.0        NaN
0.0
61166         1        NaN        0    0.0         NaN        NaN
0.0
24462         0        NaN        0    0.0         NaN        NaN
1.0

        Pioneer  Outgoing  Inspiring  Beautiful  Fun  Authentic
Credible  \
109843      NaN       0.0        1.0          1    0          1
1
28247       NaN       0.0        0.0          1    0          0
0
104792      NaN       0.0        1.0          0    0          0
0
57761       NaN       NaN        NaN          1    1          0
0
20595       NaN       1.0        1.0          0    1          0
1
62408       NaN       0.0        0.0          0    0          0
0
6553        NaN       0.0        1.0          1    0          1
0
```

```
105115      NaN       0.0        1.0          0    0          0
1
107250      0.0       NaN        NaN          0    1          0
0
31728       NaN       NaN        NaN          0    1          0
0
72534       NaN       0.0        0.0          0    1          0
0
30895       NaN       NaN        NaN          1    1          1
1
75007       NaN       0.0        0.0          0    0          0
0
61166       NaN       0.0        0.0          1    1          0
0
24462       NaN       0.0        0.0          1    1          0
0

        Way out  Cool  Catchy  Sensitive  Mainstream  Superficial
Annoying  \
109843      NaN     1     1.0        0.0         NaN          0.0
NaN
28247       NaN     0     1.0        0.0         NaN          0.0
NaN
104792      NaN     0     0.0        0.0         NaN          0.0
NaN
57761       0.0     0     0.0        NaN         NaN          NaN
NaN
20595       NaN     1     1.0        0.0         NaN          0.0
NaN
62408       NaN     1     1.0        0.0         NaN          0.0
NaN
6553        NaN     1     1.0        1.0         NaN          0.0
NaN
105115      NaN     0     1.0        0.0         NaN          0.0
NaN
107250      0.0     0     NaN        NaN         NaN          NaN
NaN
31728       0.0     0     1.0        NaN         NaN          NaN
NaN
72534       NaN     0     0.0        0.0         0.0          0.0
0.0
30895       0.0     1     1.0        NaN         NaN          NaN
NaN
75007       NaN     0     1.0        0.0         0.0          0.0
0.0
61166       NaN     1     0.0        0.0         1.0          0.0
NaN
24462       NaN     0     1.0        0.0         NaN          0.0
NaN
```

```
        Dark  Passionate  Not authentic  Good Lyrics  Background
Timeless  \
109843   NaN           1            NaN          NaN         NaN
1
28247    NaN           1            NaN          NaN         NaN
0
104792   NaN           0            NaN          NaN         NaN
1
57761    NaN           0            NaN          0.0         0.0
0
20595    NaN           1            NaN          NaN         NaN
0
62408    NaN           1            NaN          NaN         NaN
0
6553     NaN           1            NaN          NaN         NaN
1
105115   NaN           0            NaN          NaN         NaN
1
107250   0.0           1            NaN          0.0         0.0
0
31728    NaN           1            NaN          1.0         0.0
0
72534    NaN           0            0.0          NaN         NaN
0
30895    NaN           0            NaN          1.0         0.0
0
75007    NaN           0            0.0          NaN         NaN
0
61166    NaN           0            NaN          NaN         NaN
0
24462    NaN           0            NaN          NaN         NaN
0

        Depressing  Original  Talented  Worldly  Distinctive
Approachable  \
109843         0.0         1         0      NaN            0
0
28247          0.0         1         0      NaN            0
0
104792         0.0         0         1      NaN            0
0
57761          NaN         0         0      NaN            1
0
20595          0.0         0         1      NaN            0
0
62408          0.0         0         0      NaN            0
0
6553           1.0         1         1      NaN            1
1
105115         0.0         1         1      NaN            1
```

```
1
107250         NaN         1         1      0.0            0
0
31728          NaN         0         1      NaN            0
0
72534          0.0         0         0      NaN            0
0
30895          NaN         1         1      NaN            1
0
75007          0.0         0         0      NaN            0
0
61166          0.0         0         0      NaN            0
0
24462          0.0         1         0      NaN            0
0

        Genius  Trendsetter  Noisy  Upbeat  Relatable  Energetic
Exciting  \
109843     NaN            0    0.0     0.0        NaN          0
NaN
28247      NaN            0    0.0     0.0        NaN          0
NaN
104792     NaN            0    0.0     0.0        NaN          0
NaN
57761      0.0            0    NaN     1.0        NaN          1
1.0
20595      NaN            0    0.0     0.0        NaN          1
NaN
62408      NaN            0    0.0     0.0        NaN          1
NaN
6553       NaN            0    0.0     0.0        NaN          1
NaN
105115     NaN            0    0.0     0.0        NaN          1
NaN
107250     0.0            0    NaN     NaN        NaN          0
1.0
31728      0.0            0    NaN     0.0        NaN          0
0.0
72534      NaN            0    0.0     0.0        0.0          0
NaN
30895      0.0            0    NaN     0.0        NaN          1
1.0
75007      NaN            0    0.0     0.0        0.0          0
NaN
61166      NaN            0    1.0     0.0        0.0          1
NaN
24462      NaN            0    0.0     0.0        NaN          0
NaN

        Emotional  Nostalgic  None of these  Progressive  Sexy
```

```
Over  \
109843        NaN        NaN               0          NaN     0   0.0
28247         NaN        NaN               0          NaN     0   0.0
104792        NaN        NaN               0          NaN     0   0.0
57761         0.0        NaN               0          NaN     0   NaN
20595         NaN        NaN               0          NaN     1   NaN
62408         NaN        NaN               0          NaN     0   0.0
6553          NaN        NaN               0          NaN     0   0.0
105115        NaN        NaN               0          NaN     0   0.0
107250        0.0        0.0               0          0.0     1   NaN
31728         0.0        NaN               0          NaN     0   NaN
72534         NaN        NaN               0          NaN     0   0.0
30895         1.0        NaN               0          NaN     0   NaN
75007         NaN        NaN               0          NaN     0   0.0
61166         NaN        NaN               0          NaN     0   0.0
24462         NaN        NaN               0          NaN     0   0.0


       Rebellious  Fake  Cheesy  Popular  Superstar  Relaxed
Intrusive  \
109843         NaN   0.0     0.0      NaN        NaN      NaN
NaN
28247          NaN   0.0     0.0      NaN        NaN      NaN
NaN
104792         NaN   0.0     0.0      NaN        NaN      NaN
NaN
57761          0.0   NaN     NaN      0.0        NaN      0.0
NaN
20595          NaN   0.0     0.0      NaN        NaN      NaN
NaN
62408          NaN   0.0     0.0      NaN        NaN      NaN
NaN
6553           NaN   0.0     0.0      NaN        NaN      NaN
NaN
105115         NaN   0.0     0.0      NaN        NaN      NaN
```

```
NaN
107250         0.0   NaN     NaN      NaN        NaN      1.0
NaN
31728          0.0   NaN     NaN      1.0        NaN      0.0
NaN
72534          NaN   0.0     1.0      NaN        0.0      NaN
0.0
30895          0.0   NaN     NaN      1.0        NaN      0.0
NaN
75007          NaN   0.0     0.0      NaN        0.0      NaN
0.0
61166          NaN   0.0     0.0      NaN        0.0      NaN
NaN
24462          NaN   0.0     0.0      NaN        NaN      NaN
NaN

        Unoriginal  Dated  Iconic  Unapproachable  Classic  Playful
Arrogant  \
109843         0.0    0.0     NaN             0.0      1.0      1.0
0.0
28247          0.0    0.0     NaN             0.0      1.0      0.0
0.0
104792         0.0    0.0     NaN             0.0      1.0      0.0
0.0
57761          NaN    0.0     NaN             NaN      0.0      NaN
NaN
20595          0.0    0.0     NaN             0.0      NaN      0.0
0.0
62408          0.0    0.0     NaN             0.0      0.0      1.0
0.0
6553           0.0    0.0     NaN             0.0      1.0      0.0
0.0
105115         0.0    0.0     NaN             0.0      1.0      0.0
0.0
107250         NaN    NaN     0.0             NaN      0.0      NaN
NaN
31728          NaN    0.0     NaN             NaN      0.0      NaN
NaN
72534          0.0    0.0     NaN             1.0      0.0      0.0
0.0
30895          NaN    0.0     NaN             NaN      0.0      NaN
NaN
75007          0.0    0.0     NaN             0.0      0.0      0.0
0.0
61166          0.0    0.0     NaN             0.0      1.0      0.0
0.0
24462          0.0    0.0     NaN             0.0      1.0      0.0
0.0

        Warm  Soulful  Unnamed: 87  plus_score  minus_score
```

```
words_score
109843     0      NaN          NaN        12.0          0.0
12.0
28247      0      NaN          NaN         6.0          0.0
6.0
104792     0      NaN          NaN         4.0          0.0
4.0
57761      0      0.0          NaN        11.0          0.0
11.0
20595      0      NaN          NaN        14.0          0.0
14.0
62408      0      NaN          NaN         6.0          0.0
6.0
6553       0      NaN          NaN        17.0          1.0
16.0
105115     0      NaN          NaN        12.0          0.0
12.0
107250     1      NaN          NaN        10.0          0.0
10.0
31728      0      0.0          NaN         6.0          0.0
6.0
72534      0      NaN          NaN         1.0          2.0         -
1.0
30895      1      1.0          NaN        23.0          0.0
23.0
75007      0      NaN          NaN         1.0          0.0
1.0
61166      0      NaN          NaN         6.0          1.0
5.0
24462      1      NaN          NaN         8.0          0.0
8.0
```

As now we gave the word score we don't need the words columns in the words_df dataframe. Now we will create a dateframe where the columns will be the **word score of above 90**

```
words_red_df = words_df[['Artist', 'User', 'HEARD_OF',
'OWN_ARTIST_MUSIC', 'LIKE_ARTIST', 'words_score']]
```

```
words_red_df
```

```
        Artist   User                                  HEARD_OF  \
0           47  45969                                  Heard of
1           35  29118                         Never heard of
2           14  31544                                  Heard of
3           23  18085                         Never heard of
4           23  18084                         Never heard of
...        ...    ...                                       ...
118296       4   3932      Heard of and listened to music EVER
118297       4   3935      Heard of and listened to music EVER
118298      12  11216  Heard of and listened to music RECENTLY
```

```
118299      33  35142      Heard of and listened to music EVER
118300       4   3915      Heard of and listened to music EVER

                  OWN_ARTIST_MUSIC  LIKE_ARTIST  words_score
0                              NaN          NaN         -1.0
1                              NaN          NaN          3.0
2                              NaN          NaN          2.0
3                              NaN          NaN         -1.0
4                              NaN          NaN          0.0
...                            ...          ...          ...
118296  Own a little of their music         26.0         -1.0
118297  Own a little of their music         30.0          1.0
118298      Own none of their music         71.0          6.0
118299      Own none of their music         31.0          3.0
118300  Own a little of their music         46.0          4.0

[118301 rows x 6 columns]
```

```
words_red_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 118301 entries, 0 to 118300
Data columns (total 6 columns):
 #   Column            Non-Null Count   Dtype
---  ------            --------------   -----
 0   Artist            118301 non-null  int64
 1   User              118301 non-null  int64
 2   HEARD_OF          118277 non-null  object
 3   OWN_ARTIST_MUSIC  33507 non-null   object
 4   LIKE_ARTIST       33308 non-null   float64
 5   words_score       118301 non-null  float64
dtypes: float64(2), int64(2), object(2)
memory usage: 5.4+ MB
```

Merging

Now we will merge words_red_df & users_df into training_merge_df dataframe

```
users_df.rename(columns={'RESPID': 'User'}, inplace=True)
```

```
training_merge_df = train_df.merge(words_red_df, how='left',
on=['Artist', 'User'])
```

```
users_df
```

```
        User  GENDER   AGE                                WORKING
REGION  \
0      36927  Female  60.0                                  Other
South
1       3566  Female  36.0  Full-time housewife / househusband
```

```
South
2      20054  Female  52.0           Employed 30+ hours a week
Midlands
3      41749  Female  40.0        Employed 8-29 hours per week
South
4      23108  Female  16.0                   Full-time student
North
...      ...     ...   ...                                 ...       .
..
48640  19361    Male  48.0                       Self-employed
Midlands
48641  17639  Female  60.0  Full-time housewife / househusband
Midlands
48642  28753  Female  25.0           Employed 30+ hours a week
Midlands
48643  26197    Male  44.0           Employed 30+ hours a week
Midlands
48644  16225  Female  43.0                                 NaN
North

                                                    MUSIC
LIST_OWN  \
0      Music is important to me but not necessarily m...             1
hour
1      Music is important to me but not necessarily m...             1
hour
2      I like music but it does not feature heavily i...             1
hour
3       Music means a lot to me and is a passion of mine            2
hours
4       Music means a lot to me and is a passion of mine            3
hours
...                                                  ...
...
48640  I like music but it does not feature heavily i...  Less than an
hour
48641   Music means a lot to me and is a passion of mine            2
hours
48642   Music means a lot to me and is a passion of mine            2
hours
48643   Music means a lot to me and is a passion of mine            2
hours
48644  I like music but it does not feature heavily i...
NaN

               LIST_BACK    Q1    Q2    Q3    Q4    Q5    Q6     Q7
Q8  \
0                    NaN  49.0  50.0  49.0  50.0  32.0  33.0   32.0
0.0
1                 1 hour  55.0  55.0  62.0   9.0   9.0   9.0   10.0
```

```
11.0
2      Less than an hour  11.0  50.0   9.0   8.0  45.0  10.0   30.0
29.0
3                3 hours  81.0  80.0  88.0  88.0  31.0  31.0   51.0
30.0
4                6 hours  76.0  79.0  78.0  73.0  71.0  68.0   73.0
67.0
...                  ...   ...   ...   ...   ...   ...   ...    ...
...
48640            2 hours   9.0  73.0  33.0   6.0  10.0  68.0   51.0
52.0
48641             1 hour  26.0  50.0  49.0  58.0  59.0  48.0    6.0
5.0
48642            6 hours  89.0  89.0  89.0   6.0   6.0  51.0   26.0
5.0
48643            4 hours  95.0  97.0  97.0  98.0  97.0  99.0  100.0
99.0
48644                  2  49.0  48.0  50.0  51.0  49.0  25.0   45.0
51.0

          Q9   Q10   Q11   Q12   Q13   Q14   Q15   Q16    Q17   Q18
Q19
0       74.0  50.0  50.0  71.0  52.0  71.0   9.0   7.0   72.0  49.0
26.0
1       55.0  12.0  65.0  65.0  80.0  79.0  51.0  31.0   68.0  54.0
33.0
2        8.0  50.0  94.0  51.0  74.0  66.0  27.0  46.0   73.0   8.0
31.0
3        8.0  76.0  74.0  64.0  73.0  85.0  61.0  77.0   76.0  78.0
88.0
4       31.0  56.0  13.0  82.0  79.0  68.0  71.0   NaN   86.0  80.0
32.0
...      ...   ...   ...   ...   ...   ...   ...   ...    ...   ...
...
48640   93.0  53.0  74.0  36.0  13.0  38.0  12.0  10.0   50.0  10.0
28.0
48641   88.0  58.0  62.0  79.0  17.0  24.0  30.0   6.0   73.0  20.0
21.0
48642    0.0  70.0  70.0  70.0  51.0  70.0  70.0   NaN  100.0  70.0
69.0
48643  100.0  97.0  98.0  99.0  97.0  99.0  99.0  99.0  100.0  91.0
96.0
48644   40.0  10.0  69.0  70.0  53.0  54.0  10.0   4.0    7.0   NaN
NaN

[48645 rows x 27 columns]
```

training_merge_df

```
        Artist  Track   User  Rating  Time  \
0           40    179  47994       9    17
1            9     23   8575      58     7
2           46    168  45475      13    16
3           11    153  39508      42    15
4           14     32  11565      54    19
...        ...    ...    ...     ...   ...
188685       0      3   1278      29     6
188686       1      6   2839      30    18
188687      10    142  35756      61    12
188688      22     54  20163      46    21
188689      47    171  45580      12     4

                                       HEARD_OF
OWN_ARTIST_MUSIC  \
0                                 Never heard of
NaN
1                                 Never heard of
NaN
2                                 Never heard of
NaN
3           Heard of and listened to music EVER   Own none of their
music
4           Heard of and listened to music EVER   Own none of their
music
...                                          ...
...
188685                            Never heard of
NaN
188686                                 Heard of
NaN
188687                                 Heard of
NaN
188688  Heard of and listened to music RECENTLY  Own a lot of their
music
188689  Heard of and listened to music RECENTLY   Own none of their
music

        LIKE_ARTIST  words_score
0               NaN         -2.0
1               NaN          5.0
2               NaN          1.0
3              28.0          4.0
4              18.0          2.0
...             ...          ...
188685          NaN          3.0
188686          NaN         -1.0
188687          NaN          3.0
188688         74.0         10.0
188689          7.0          1.0
```

```
[188690 rows x 9 columns]
```

```
training_merge_df = training_merge_df.merge(users_df, how='left',
on=['User'])
```

```
training_merge_df
```

```
        Artist  Track   User  Rating  Time  \
0           40    179  47994       9    17
1            9     23   8575      58     7
2           46    168  45475      13    16
3           11    153  39508      42    15
4           14     32  11565      54    19
...        ...    ...    ...     ...   ...
188685       0      3   1278      29     6
188686       1      6   2839      30    18
188687      10    142  35756      61    12
188688      22     54  20163      46    21
188689      47    171  45580      12     4

                                       HEARD_OF
OWN_ARTIST_MUSIC  \
0                                Never heard of
NaN
1                                Never heard of
NaN
2                                Never heard of
NaN
3          Heard of and listened to music EVER   Own none of their
music
4          Heard of and listened to music EVER   Own none of their
music
...                                         ...
...
188685                           Never heard of
NaN
188686                                Heard of
NaN
188687                                Heard of
NaN
188688  Heard of and listened to music RECENTLY  Own a lot of their
music
188689  Heard of and listened to music RECENTLY   Own none of their
music

        LIKE_ARTIST  words_score  GENDER   AGE  \
0               NaN         -2.0  Female  41.0
1               NaN          5.0  Female  45.0
2               NaN          1.0    Male  23.0
```

```
3              28.0          4.0  Female  61.0
4              18.0          2.0  Female  20.0
...             ...          ...     ...   ...
188685          NaN          3.0  Female  53.0
188686          NaN         -1.0    Male  52.0
188687          NaN          3.0  Female  28.0
188688         74.0         10.0  Female  35.0
188689          7.0          1.0  Female  82.0

                                    WORKING    REGION  \
0                    Temporarily unemployed     North
1                                       NaN    Centre
2              Employed 8-29 hours per week  Midlands
3              Retired from self-employment  Midlands
4                    Temporarily unemployed     South
...                                      ...       ...
188685                                  NaN     North
188686            Employed 30+ hours a week  Midlands
188687  Full-time housewife / househusband     North
188688            Employed 30+ hours a week     North
188689                                  NaN    Centre

                                                      MUSIC
LIST_OWN  \
0        Music means a lot to me and is a passion of mine            3
hours
1       Music is important to me but not necessarily m...
1
2        Music means a lot to me and is a passion of mine            5
hours
3       Music is important to me but not necessarily m...
1 hour
4       Music is important to me but not necessarily m...  Less than
an hour
...                                                   ...
...
188685  Music is important to me but not necessarily m...
1
188686  I like music but it does not feature heavily i...
1 hour
188687  Music is important to me but not necessarily m...
NaN
188688  Music is important to me but not necessarily m...
1 hour
188689  Music is important to me but not necessarily m...
0

       LIST_BACK     Q1    Q2    Q3    Q4    Q5    Q6    Q7    Q8
Q9   Q10  \
0         0 Hours   62.0  22.0  62.0  48.0  35.0  30.0  48.0  28.0
```

```
88.0  70.0
1              2   32.0  57.0  52.0  10.0  10.0  29.0  73.0  51.0
12.0  50.0
2            NaN  100.0  75.0  90.0  48.0  25.0  34.0  46.0  29.0
29.0  71.0
3            NaN   62.0  57.0  55.0  44.0  53.0  66.0  33.0  27.0
41.0  52.0
4        3 hours   22.0  69.0  28.0  52.0  32.0  22.0   9.0  10.0
11.0  55.0
...          ...    ...   ...   ...   ...   ...   ...   ...   ...   ..
.   ...
188685       NaN   68.0  52.0  66.0  49.0  49.0  31.0  30.0   8.0
29.0  49.0
188686    1 hour   75.0  50.0  32.0   7.0  48.0  50.0  66.0  30.0
48.0  48.0
188687       NaN   52.0  67.0  51.0  52.0  53.0  35.0  15.0  14.0
53.0  51.0
188688    1 hour    9.0  27.0  13.0  13.0   6.0  58.0  45.0  30.0
61.0  13.0
188689         0   73.0  92.0  93.0   7.0  11.0   9.0  11.0   9.0
34.0  73.0

         Q11   Q12    Q13    Q14   Q15   Q16   Q17   Q18   Q19
0       49.0  49.0   32.0   32.0  50.0  31.0  31.0  10.0   9.0
1       91.0  72.0   32.0   55.0  53.0  54.0  75.0   NaN   NaN
2       72.0  48.0  100.0  100.0  28.0  65.0  72.0  73.0  83.0
3       71.0  73.0   53.0   61.0  49.0  52.0  63.0  50.0  45.0
4       84.0  70.0   20.0   19.0  11.0  47.0  71.0  37.0  26.0
...      ...   ...    ...    ...   ...   ...   ...   ...   ...
188685  74.0  69.0   50.0   30.0  11.0  51.0  51.0   NaN   NaN
188686  48.0  30.0   30.0   49.0  32.0  32.0  47.0  31.0   8.0
188687  50.0  51.0   57.0   51.0  52.0  52.0  52.0  54.0  47.0
188688  54.0  49.0   65.0   50.0   4.0  46.0  77.0  47.0  39.0
188689  73.0  73.0   54.0   69.0   8.0  10.0  70.0   NaN   NaN

[188690 rows x 35 columns]

training_merge_df.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 188690 entries, 0 to 188689
Data columns (total 35 columns):
 #   Column            Non-Null Count   Dtype
---  ------            --------------   -----
 0   Artist            188690 non-null  int64
 1   Track             188690 non-null  int64
 2   User              188690 non-null  int64
 3   Rating            188690 non-null  int64
 4   Time              188690 non-null  int64
 5   HEARD_OF          186418 non-null  object
```

```
 6   OWN_ARTIST_MUSIC  56835 non-null   object
 7   LIKE_ARTIST       55028 non-null   float64
 8   words_score       186636 non-null  float64
 9   GENDER            176833 non-null  object
 10  AGE               174982 non-null  float64
 11  WORKING           140545 non-null  object
 12  REGION            167481 non-null  object
 13  MUSIC             176833 non-null  object
 14  LIST_OWN          158651 non-null  object
 15  LIST_BACK         158790 non-null  object
 16  Q1                176833 non-null  float64
 17  Q2                176833 non-null  float64
 18  Q3                176833 non-null  float64
 19  Q4                176833 non-null  float64
 20  Q5                176833 non-null  float64
 21  Q6                176833 non-null  float64
 22  Q7                176833 non-null  float64
 23  Q8                176833 non-null  float64
 24  Q9                176833 non-null  float64
 25  Q10               176833 non-null  float64
 26  Q11               176833 non-null  float64
 27  Q12               176833 non-null  float64
 28  Q13               176833 non-null  float64
 29  Q14               176833 non-null  float64
 30  Q15               176833 non-null  float64
 31  Q16               142754 non-null  float64
 32  Q17               176833 non-null  float64
 33  Q18               140545 non-null  float64
 34  Q19               140545 non-null  float64
dtypes: float64(22), int64(5), object(8)
memory usage: 51.8+ MB

training_merge_df.sample(15)

        Artist  Track   User  Rating  Time  \
52030       35     91  30678      58    23
175760       4     12   5536      89    18
164836      11     29   9928      12     7
101302      34     87  28862      41    23
151739      40    176  50514       3    17
96733       30     77  22275      10    22
87059       46    169  45280      10    16
100404      22    128  32931      10     0
59518       40    176  50861      64    17
65877       22    114  32187      32     0
142221      22    122  33185      24     0
31295        6     14   5817      50     7
164837      15     33  13203       5    19
40160        2     69  21549      66    22
20632       48    172  47979      56    17
```

```
                                      HEARD_OF
OWN_ARTIST_MUSIC  \
52030                            Never heard of
NaN
175760      Heard of and listened to music EVER     Own a lot of their
music
164836                           Never heard of
NaN
101302                           Never heard of
NaN
151739                           Never heard of
NaN
96733                            Never heard of
NaN
87059                                 Heard of
NaN
100404                      Ever heard music by      Own none of their
music
59518   Heard of and listened to music RECENTLY     Own a lot of their
music
65877                       Ever heard music by  Own a little of their
music
142221                      Ever heard music by      Own none of their
music
31295       Heard of and listened to music EVER  Own a little of their
music
164837                           Never heard of
NaN
40160                            Never heard of
NaN
20632                            Never heard of
NaN

        LIKE_ARTIST  words_score  GENDER   AGE  \
52030           NaN          6.0  Female  29.0
175760         72.0          6.0  Female  60.0
164836          NaN          0.0    Male  53.0
101302          NaN          2.0  Female  18.0
151739          NaN         -1.0    Male  40.0
96733           NaN         -2.0     NaN   NaN
87059           NaN         -1.0    Male  64.0
100404          8.0          0.0  Female  34.0
59518          87.0          3.0  Female  29.0
65877          30.0          8.0    Male  30.0
142221         27.0          4.0  Female  24.0
31295          50.0         12.0    Male  18.0
164837          NaN         -3.0    Male  40.0
40160           NaN          4.0     NaN   NaN
20632           NaN          2.0  Female  43.0
```

```
                                                      WORKING
REGION  \
52030                           Employed 30+ hours a week
Midlands
175760  Retired from full-time employment (30+ hours p...
South
164836                                                NaN
South
101302      Employed part-time less than 8 hours per week
South
151739                                       Self-employed  Northern
Ireland
96733                                                 NaN
NaN
87059                                               Other
North
100404                                                NaN
NaN
59518                           Employed 30+ hours a week
Midlands
65877                                                 NaN
NaN
142221                                                NaN
NaN
31295                                                 NaN
North
164837                                       Self-employed
South
40160                                                 NaN
NaN
20632                           Employed 30+ hours a week
Midlands

                                                      MUSIC
LIST_OWN  \
52030   Music is important to me but not necessarily m...    2 hours

175760  Music is no longer as important as it used to ...    0 Hours

164836  I like music but it does not feature heavily i...        NaN

101302  Music is important to me but not necessarily m...    8 hours

151739   Music means a lot to me and is a passion of mine  16+ hours

96733                                                 NaN        NaN

87059    Music means a lot to me and is a passion of mine    2 hours
```

```
100404   Music means a lot to me and is a passion of mine          2
59518   Music is important to me but not necessarily m...     1 hour
65877   Music is important to me but not necessarily m...          5
142221  Music is important to me but not necessarily m...        NaN
31295   Music is important to me but not necessarily m...          3
164837  Music is no longer as important as it used to ...    0 Hours
40160                                                  NaN        NaN
20632   Music is important to me but not necessarily m...     1 hour


       LIST_BACK     Q1     Q2     Q3    Q4    Q5     Q6     Q7     Q8
Q9  \
52030    5 hours   62.0   68.0   83.0  67.0  56.0    1.0    1.0    1.0
54.0
175760   2 hours   51.0   52.0   51.0  52.0  70.0   52.0   53.0   54.0
74.0
164836       NaN    2.0  100.0    2.0   2.0   2.0   99.0    2.0    2.0
100.0
101302    1 hour   73.0   78.0   88.0  95.0  99.0   74.0   93.0   82.0
15.0
151739   0 Hours  100.0  100.0  100.0   3.0  50.0  100.0   20.0   27.0
39.0
96733        NaN    NaN    NaN    NaN   NaN   NaN    NaN    NaN    NaN
NaN
87059    4 hours   72.0   72.0   72.0  31.0  29.0   66.0   20.0   20.0
19.0
100404         4   41.0   79.0   85.0  30.0  35.0    9.0   12.0   15.0
13.0
59518    3 hours    8.0   71.0   49.0  13.0  30.0   14.0   69.0   64.0
28.0
65877          7   67.0   96.0   82.0  14.0  14.0   13.0    4.0    5.0
28.0
142221         9   67.0   64.0   66.0  43.0  60.0   27.0   67.0   77.0
18.0
31295        NaN  100.0  100.0   65.0  75.0  72.0    9.0  100.0  100.0
12.0
164837  10 hours   16.0   48.0    9.0  19.0  50.0   51.0   37.0   31.0
58.0
40160        NaN    NaN    NaN    NaN   NaN   NaN    NaN    NaN    NaN
NaN
20632     1 hour   34.0   65.0   46.0  18.0  40.0    9.0   10.0   12.0
```

```
32.0

          Q10    Q11    Q12    Q13   Q14   Q15   Q16    Q17   Q18
Q19
52030    85.0   66.0   49.0   91.0  74.0  55.0   NaN   73.0  73.0
70.0
175760   38.0   75.0   47.0   48.0  49.0  39.0  53.0   55.0  53.0
53.0
164836    2.0    2.0    2.0    2.0   2.0   2.0   2.0    2.0   NaN
NaN
101302   94.0   90.0   90.0   74.0  81.0  72.0   NaN   86.0  81.0
96.0
151739   49.0   33.0   40.0   36.0  78.0  50.0   3.0    3.0  50.0
25.0
96733     NaN    NaN    NaN    NaN   NaN   NaN   NaN    NaN   NaN
NaN
87059    54.0   17.0   18.0   23.0  69.0  69.0   8.0   10.0  11.0
11.0
100404   84.0   15.0   37.0   88.0  83.0  89.0  76.0   43.0   NaN
NaN
59518    46.0   66.0   46.0   31.0  43.0  29.0  29.0   47.0  48.0
47.0
65877    78.0   99.0   81.0    7.0  28.0  76.0  38.0   94.0   NaN
NaN
142221   61.0   81.0   71.0   50.0  61.0  60.0  56.0   61.0   NaN
NaN
31295   100.0  100.0  100.0  100.0  68.0  48.0   0.0  100.0   NaN
NaN
164837   38.0   47.0   47.0   33.0  30.0  11.0  12.0   15.0  16.0
8.0
40160     NaN    NaN    NaN    NaN   NaN   NaN   NaN    NaN   NaN
NaN
20632    59.0   66.0   66.0   94.0  62.0  31.0  29.0   64.0  40.0
30.0
```

Merging the test dataset

```
test_merge_df = test_df.merge(words_red_df, how='left', on=['Artist',
'User'])
test_merge_df = test_merge_df.merge(users_df, how='left', on=['User'])

test_merge_df
```

```
        Artist  Track   User  Time
HEARD_OF  \
0            1      6   3475    18  Heard of and listened to music
EVER
1            6    149  39210    15
NaN
2           40    177  47861    17                       Never heard
of
```

```
3           31     79  27413    11                       Never heard
of
4           26     66  23232    22                       Never heard
of
...        ...    ...    ...   ...                                ..
.
125789      14     95  30004    23                            Heard
of
125790      10     25   8186     7                       Never heard
of
125791      40    146  38180    13                            Heard
of
125792      22    113  32918     0                  Ever heard music
by
125793       2     70  24231    22                       Never heard
of

                OWN_ARTIST_MUSIC  LIKE_ARTIST  words_score  GENDER
AGE  \
0       Own none of their music          3.0          2.0  Female
48.0
1                           NaN          NaN          NaN    Male
28.0
2                           NaN          NaN         -2.0  Female
59.0
3                           NaN          NaN          0.0  Female
25.0
4                           NaN          NaN          0.0     NaN
NaN
...                         ...          ...          ...     ...   ..
.
125789                      NaN          NaN         12.0    Male
36.0
125790                      NaN          NaN          6.0    Male
49.0
125791                      NaN          NaN          3.0  Female
40.0
125792  Own none of their music         48.0          2.0  Female
48.0
125793                      NaN          NaN          4.0    Male
43.0

                                                 WORKING    REGION  \
0                           Employed 30+ hours a week     South
1                           Employed 30+ hours a week  Midlands
2                                               Other  Midlands
3       Employed part-time less than 8 hours per week  Midlands
4                                                 NaN       NaN
...                                               ...       ...
125789                      Employed 30+ hours a week  Midlands
```

```
125790                                          NaN     North
125791            Full-time housewife / househusband  Midlands
125792                                          NaN       NaN
125793                     Employed 30+ hours a week     North

                                                  MUSIC
LIST_OWN  \
0        Music means a lot to me and is a passion of mine
1 hour
1       Music is important to me but not necessarily m...
1 hour
2       Music is no longer as important as it used to ...  Less than
an hour
3       I like music but it does not feature heavily i...
1 hour
4                                                    NaN
NaN
...                                                  ...
...
125789   Music means a lot to me and is a passion of mine           10
hours
125790  I like music but it does not feature heavily i...
NaN
125791   Music means a lot to me and is a passion of mine           15
hours
125792   Music means a lot to me and is a passion of mine
0
125793  Music is important to me but not necessarily m...
1 hour

               LIST_BACK    Q1    Q2     Q3    Q4    Q5    Q6    Q7
Q8  \
0                3 hours   8.0  69.0   27.0  27.0  50.0  27.0  26.0
8.0
1                 1 hour  81.0  67.0   94.0  61.0  53.0  32.0  41.0
42.0
2       Less than an hour   9.0  94.0   49.0  48.0  49.0   8.0  13.0
56.0
3                 1 hour  53.0  38.0   51.0  53.0  53.0  53.0  33.0
51.0
4                    NaN   NaN   NaN    NaN   NaN   NaN   NaN   NaN
NaN
...                  ...   ...   ...    ...   ...   ...   ...   ...
...
125789           4 hours  84.0  69.0  100.0  32.0   9.0  28.0   9.0
12.0
125790                 3  29.0  70.0   30.0  30.0  69.0  14.0  12.0
12.0
125791           9 hours  59.0  51.0   51.0  83.0  32.0  43.0  14.0
41.0
```

```
125792                  1  69.0  30.0   76.0  74.0  73.0  11.0  11.0
11.0
125793            3 hours  15.0  68.0   51.0  51.0  51.0  71.0   2.0
2.0

           Q9    Q10   Q11   Q12   Q13   Q14   Q15   Q16   Q17   Q18
Q19
0        51.0  50.0  66.0  49.0  20.0   7.0   8.0   9.0   7.0   4.0
8.0
1        36.0  76.0  70.0  76.0  58.0  61.0  66.0  51.0  75.0  70.0
72.0
2        92.0  92.0  55.0  57.0  11.0  57.0  10.0  11.0  91.0   7.0
9.0
3        47.0  33.0  41.0  45.0  49.0  49.0  49.0  49.0  35.0  52.0
52.0
4         NaN   NaN   NaN   NaN   NaN   NaN   NaN   NaN   NaN   NaN
NaN
...       ...   ...   ...   ...   ...   ...   ...   ...   ...   ...   .
..
125789   50.0  75.0  68.0  72.0  64.0  70.0  75.0   NaN  72.0  56.0
54.0
125790   70.0  29.0  50.0  48.0  54.0  66.0  10.0  34.0  70.0   NaN
NaN
125791   71.0  58.0  36.0  43.0  81.0  63.0  45.0  65.0  30.0  46.0
21.0
125792   92.0  34.0  74.0  72.0  36.0  37.0   9.0   9.0  64.0   NaN
NaN
125793   94.0  65.0   2.0   3.0   3.0   3.0   3.0   NaN  30.0   5.0
5.0

[125794 rows x 34 columns]
```

Data Analysis

Now we will try to get the insights from the dataset and see if there is any relationship between the columns. We must also check if any of the columns are interdependent. We ask Question and then we visualize the dataset to get the Answer.

We can do this by plotting the graphs for various columns and observing the relation between the two or more columns depending on the plot we choose.

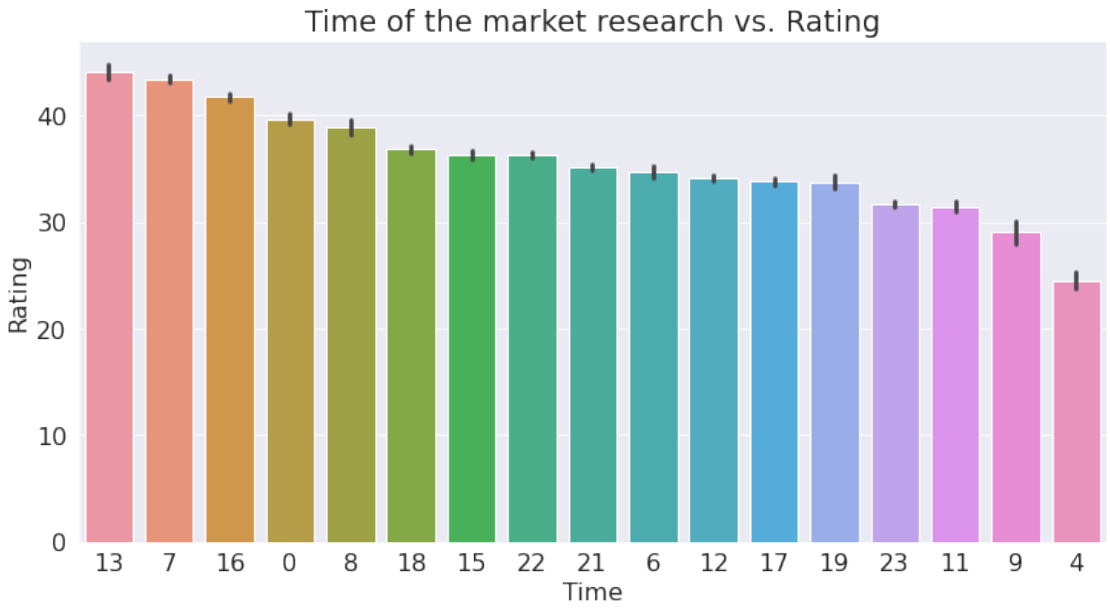
```
# Do you love or hate the the song?
px.histogram(training_merge_df, x='Rating', nbins=101, marginal='box',
title='Rating(Love/Hate) Distribution')
```

```
training_merge_df.columns
```

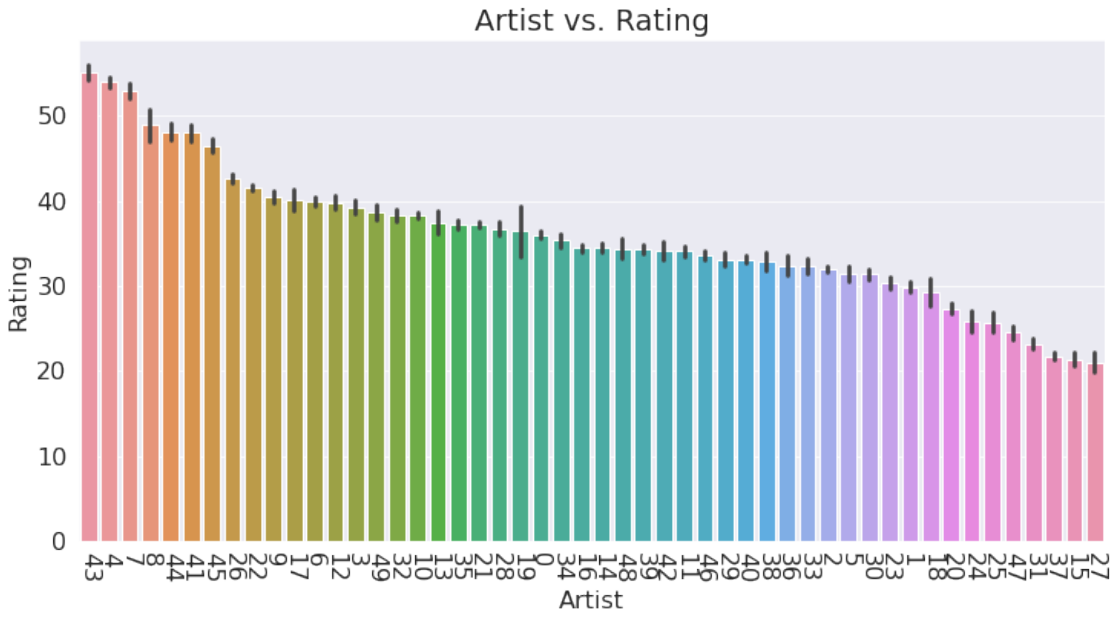
```
Index(['Artist', 'Track', 'User', 'Rating', 'Time', 'HEARD_OF',
       'OWN_ARTIST_MUSIC', 'LIKE_ARTIST', 'words_score', 'GENDER',
```

```
'AGE',
       'WORKING', 'REGION', 'MUSIC', 'LIST_OWN', 'LIST_BACK', 'Q1',
'Q2', 'Q3',
       'Q4', 'Q5', 'Q6', 'Q7', 'Q8', 'Q9', 'Q10', 'Q11', 'Q12', 'Q13',
'Q14',
       'Q15', 'Q16', 'Q17', 'Q18', 'Q19'],
      dtype='object')
```

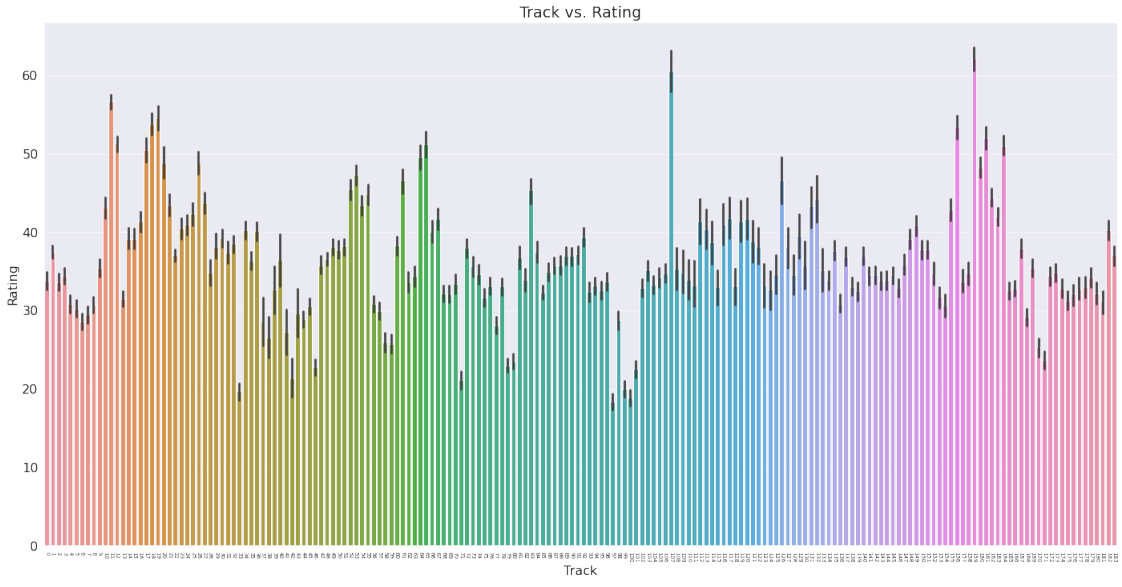
```
plot_order= training_merge_df.groupby('Time')
['Rating'].mean().sort_values(ascending=False).index.values

fig, ax = plt.subplots(figsize=(12,6))

plt.title('Time of the market research vs. Rating')
sns.barplot(x='Time', y='Rating', data=training_merge_df,
order=plot_order)
plt.xticks(rotation=0, ha='center')
plt.show();
```




```
plot_order= training_merge_df.groupby('Artist')
['Rating'].mean().sort_values(ascending=False).index.values

fig, ax = plt.subplots(figsize=(12,6))

plt.title('Artist vs. Rating')
sns.barplot(x='Artist', y='Rating', data=training_merge_df,
order=plot_order)
plt.xticks(rotation=270, ha='center')
plt.show();
```




```
fig, ax = plt.subplots(figsize=(24,12))

plt.title('Track vs. Rating')
sns.barplot(x='Track', y='Rating', data=training_merge_df)
plt.xticks(rotation=-90, fontsize=7, ha='center')
plt.show();
```



Change of columns

```
training_merge_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 188690 entries, 0 to 188689
```

```
Data columns (total 35 columns):
 #   Column             Non-Null Count   Dtype  
---  ------             --------------   -----  
 0   Artist             188690 non-null  int64  
 1   Track              188690 non-null  int64  
 2   User               188690 non-null  int64  
 3   Rating             188690 non-null  int64  
 4   Time               188690 non-null  int64  
 5   HEARD_OF           186418 non-null  object 
 6   OWN_ARTIST_MUSIC   56835 non-null   object 
 7   LIKE_ARTIST        55028 non-null   float64
 8   words_score        186636 non-null  float64
 9   GENDER             176833 non-null  object 
 10  AGE                174982 non-null  float64
 11  WORKING            140545 non-null  object 
 12  REGION             167481 non-null  object 
 13  MUSIC              176833 non-null  object 
 14  LIST_OWN           158651 non-null  object 
 15  LIST_BACK          158790 non-null  object 
 16  Q1                 176833 non-null  float64
 17  Q2                 176833 non-null  float64
 18  Q3                 176833 non-null  float64
 19  Q4                 176833 non-null  float64
 20  Q5                 176833 non-null  float64
 21  Q6                 176833 non-null  float64
 22  Q7                 176833 non-null  float64
 23  Q8                 176833 non-null  float64
 24  Q9                 176833 non-null  float64
 25  Q10                176833 non-null  float64
 26  Q11                176833 non-null  float64
 27  Q12                176833 non-null  float64
 28  Q13                176833 non-null  float64
 29  Q14                176833 non-null  float64
 30  Q15                176833 non-null  float64
 31  Q16                142754 non-null  float64
 32  Q17                176833 non-null  float64
 33  Q18                140545 non-null  float64
 34  Q19                140545 non-null  float64
dtypes: float64(22), int64(5), object(8)
memory usage: 55.9+ MB
```

```
training_merge_df['HEARD_OF'].value_counts()
```

```
Never heard of                              94090
Heard of                                    35493
Heard of and listened to music EVER         29854
Heard of and listened to music RECENTLY     17847
Ever heard music by                          5136
Listened to recently                         2191
Ever heard of                                1807
Name: HEARD_OF, dtype: int64
```

```
print('Missing values in HEARD_OF column
{}'.format(training_merge_df['HEARD_OF'].isna().sum()))
```

```
Missing values in HEARD_OF column 2272
```

```
training_merge_df['HEARD_OF'].replace(['Ever heard of'], 'Never heard
of', inplace=True)
training_merge_df['HEARD_OF'].replace(['Ever heard music by'], 'Heard
of and listened to music EVER', inplace=True)
training_merge_df['HEARD_OF'].replace(['Listened to recently'], 'Heard
of and listened to music RECENTLY', inplace=True)
training_merge_df['HEARD_OF'].fillna('Never heard of', inplace=True)
```

```
training_merge_df['HEARD_OF'].unique()
```

```
array(['Never heard of', 'Heard of and listened to music EVER',
       'Heard of', 'Heard of and listened to music RECENTLY'],
      dtype=object)
```

```
test_merge_df['HEARD_OF'].replace(['Ever heard of'], 'Never heard of',
inplace=True)
test_merge_df['HEARD_OF'].replace(['Ever heard music by'], 'Heard of
and listened to music EVER', inplace=True)
test_merge_df['HEARD_OF'].replace(['Listened to recently'], 'Heard of
and listened to music RECENTLY', inplace=True)
test_merge_df['HEARD_OF'].fillna('Never heard of', inplace=True)
```

```
test_merge_df['HEARD_OF'].unique()
```

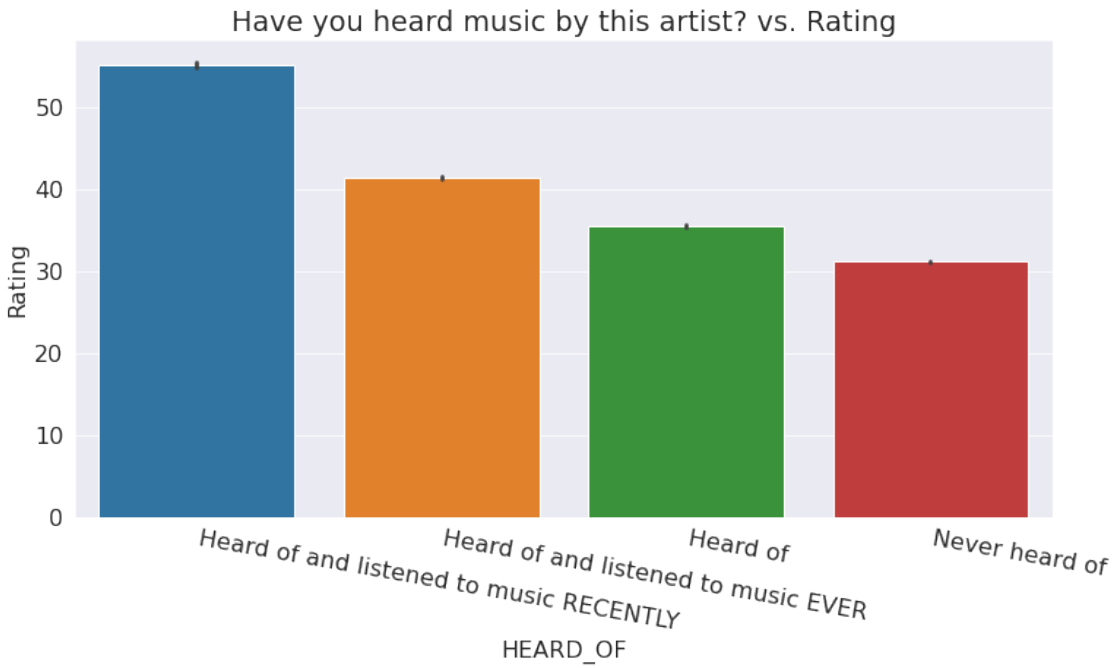
```
array(['Heard of and listened to music EVER', 'Never heard of',
       'Heard of', 'Heard of and listened to music RECENTLY'],
      dtype=object)
```

```
training_merge_df['HEARD_OF'].value_counts()
```

```
Never heard of                          98169
Heard of                                35493
Heard of and listened to music EVER     34990
Heard of and listened to music RECENTLY    20038
Name: HEARD_OF, dtype: int64
```

```
plot_order= training_merge_df.groupby('HEARD_OF')
['Rating'].mean().sort_values(ascending=False).index.values
```

```
fig, ax = plt.subplots(figsize=(12,6))

plt.title('Have you heard music by this artist? vs. Rating')
sns.barplot(x='HEARD_OF', y='Rating', data=training_merge_df,
order=plot_order)
plt.xticks(rotation=350, ha='left')
plt.show();
```



Own_Artist_Music

```
training_merge_df['OWN_ARTIST_MUSIC'].unique()
```

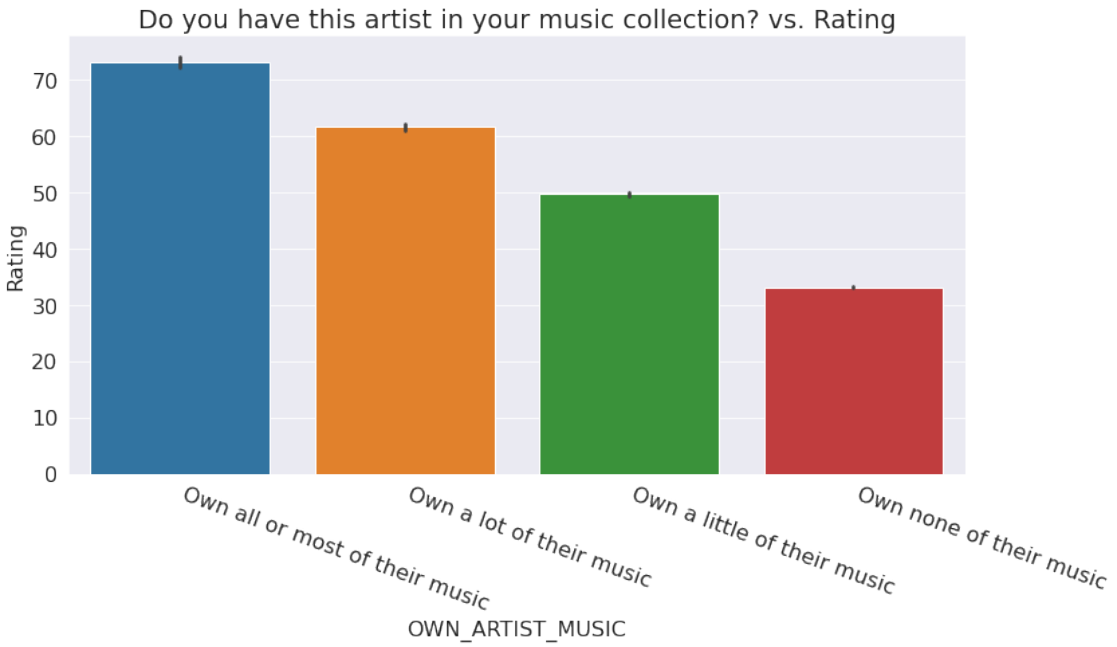
```
array([nan, 'Own none of their music', 'Own a little of their music',
       'Own all or most of their music', 'DonÕt know',
       'Own a lot of their music', 'DonÍt know', 'don`t know'],
      dtype=object)
```

```
training_merge_df['OWN_ARTIST_MUSIC'].value_counts()
```

```
Own none of their music           26810
Own a little of their music       18721
Own a lot of their music           7263
Own all or most of their music     2593
DonÕt know                         1265
DonÍt know                          147
don`t know                           36
Name: OWN_ARTIST_MUSIC, dtype: int64
```

```
training_merge_df['OWN_ARTIST_MUSIC'].replace(['DonÕt know'], 'Own
none of their music', inplace=True)
training_merge_df['OWN_ARTIST_MUSIC'].replace(['DonÍt know'], 'Own
none of their music', inplace=True)
training_merge_df['OWN_ARTIST_MUSIC'].replace(['don`t know'], 'Own
none of their music', inplace=True)
training_merge_df['OWN_ARTIST_MUSIC'].fillna('Own none of their
music', inplace=True)
```

```
test_merge_df['OWN_ARTIST_MUSIC'].replace(['DonÕt know'], 'Own none of
their music', inplace=True)
test_merge_df['OWN_ARTIST_MUSIC'].replace(['DonÍt know'], 'Own none of
their music', inplace=True)
test_merge_df['OWN_ARTIST_MUSIC'].replace(['don`t know'], 'Own none of
their music', inplace=True)
test_merge_df['OWN_ARTIST_MUSIC'].fillna('Own none of their music',
inplace=True)

training_merge_df['OWN_ARTIST_MUSIC'].unique()

array(['Own none of their music', 'Own a little of their music',
       'Own all or most of their music', 'Own a lot of their music'],
      dtype=object)

training_merge_df['OWN_ARTIST_MUSIC'].value_counts()

Own none of their music           160113
Own a little of their music        18721
Own a lot of their music            7263
Own all or most of their music      2593
Name: OWN_ARTIST_MUSIC, dtype: int64

plot_order= training_merge_df.groupby('OWN_ARTIST_MUSIC')
['Rating'].mean().sort_values(ascending=False).index.values

fig, ax = plt.subplots(figsize=(12,6))

plt.title('Do you have this artist in your music collection? vs.
Rating')
sns.barplot(x='OWN_ARTIST_MUSIC', y='Rating', data=training_merge_df,
order=plot_order)
plt.xticks(rotation=340, ha='left')
plt.show();
```



LIKE_ARTIST

```
training_merge_df['LIKE_ARTIST'].unique()
```

```
array([   nan,  28.  ,  18.  ,  33.  ,  36.  ,  53.  ,  50.  ,  63.  ,
        68.  ,  56.  ,  74.  ,  51.  ,  38.  ,  29.  ,  71.  ,  90.  ,
        70.  ,  30.  ,  52.  ,  84.  ,  59.  ,  66.  ,  42.  ,  48.  ,
        32.  ,  49.  ,  81.  , 100.  ,  45.  ,  87.  ,  57.  ,  83.  ,
        92.  ,  75.  ,  47.  ,  13.  ,  41.  ,  17.  ,  12.  ,   1.  ,
         4.  ,  55.  ,  65.  ,  16.  ,  58.  ,  99.  ,  69.  ,  15.  ,
        27.  ,  46.  ,  10.  ,  44.  ,  35.  ,   6.  ,  31.  ,  73.  ,
        26.  ,   2.  ,  43.  ,  54.  ,  61.  ,   9.  ,  14.  ,  62.  ,
        67.  ,  89.  ,  72.  ,  39.  ,   7.  ,   5.  ,  31.34,  20.  ,
        88.  ,  25.  ,  94.  ,  77.  ,  82.  ,  64.  ,  80.  ,  22.  ,
        23.  ,  86.  ,  40.  ,  37.  ,  34.  ,  21.  ,  93.  ,  11.  ,
        91.  ,  30.92,  98.  ,  79.  ,   8.  ,  33.05,   3.  ,  76.  ,
        85.  ,  78.  ,  60.  ,  24.  ,  97.  ,  19.  ,  95.  ,  29.21,
        28.14,  96.  ,  62.47,  48.83,  54.58,  23.24,  39.45,   0.  ,
        23.88,  32.84,  82.73,  78.25,  55.01,  78.68,  39.02,  65.88,
        13.01,   8.53,  38.59,  49.04,  22.6 ,  70.15,  18.34,  45.84,
        21.32,  55.44,  28.57,  37.74,  75.48,  38.17,  60.34,  32.41,
        27.51,  56.72,  80.81,  26.23,  51.81,  44.99,  57.57,  60.55,
        98.08,  16.63,  66.74,  20.9 ,  27.72,  46.91,  86.78,  62.69,
        72.92,  61.19,  29.85,  47.76,  69.72,  71.64,  84.01,  75.91,
        52.24,  29.64,  51.06,  43.28,  47.55,  25.37,   2.99,  50.32,
        80.38])
```

```
training_merge_df['LIKE_ARTIST'].value_counts()
```

```
49.00    2707
51.00    2463
30.00    2425
50.00    2218
29.00    2114
         ...
44.99       1
57.57       1
60.55       1
98.08       1
80.38       1
Name: LIKE_ARTIST, Length: 168, dtype: int64
```

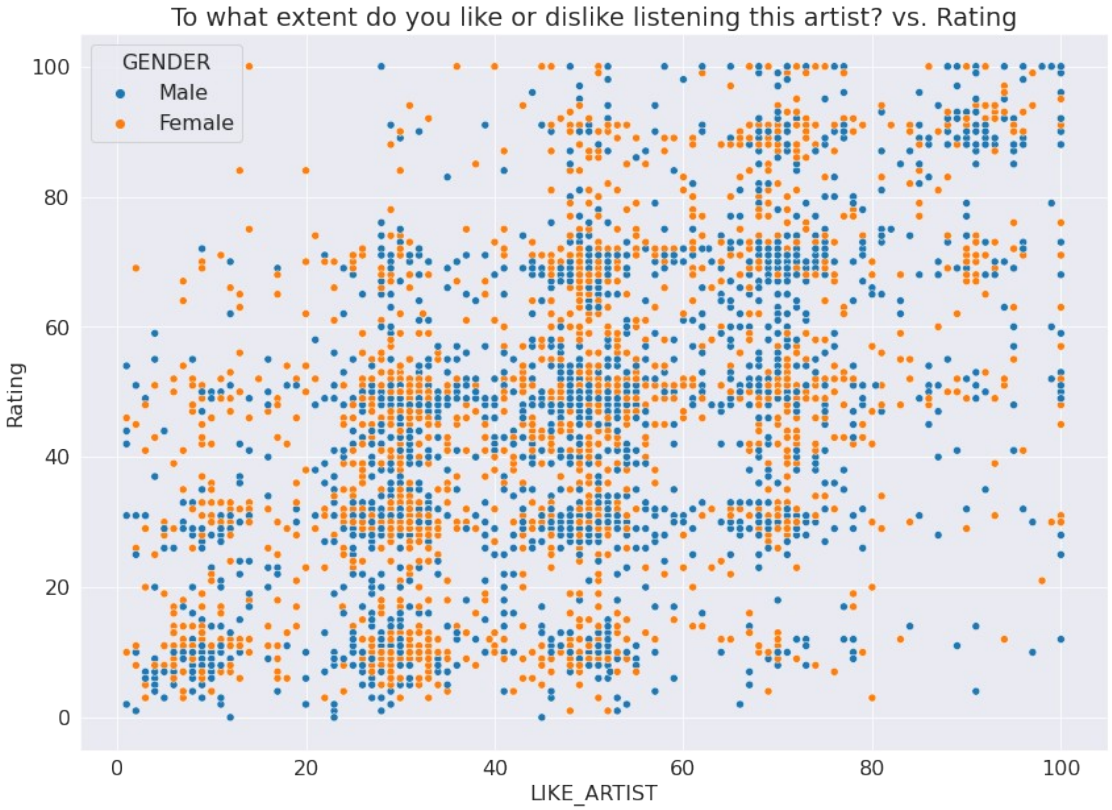
```
training_merge_df
```

```
        Artist  Track   User  Rating  Time  \
0           40    179  47994       9    17
1            9     23   8575      58     7
2           46    168  45475      13    16
3           11    153  39508      42    15
4           14     32  11565      54    19
...        ...    ...    ...     ...   ...
188685       0      3   1278      29     6
188686       1      6   2839      30    18
188687      10    142  35756      61    12
188688      22     54  20163      46    21
188689      47    171  45580      12     4

                                      HEARD_OF
OWN_ARTIST_MUSIC  \
0                                Never heard of   Own none of their
music
1                                Never heard of   Own none of their
music
2                                Never heard of   Own none of their
music
3          Heard of and listened to music EVER   Own none of their
music
4          Heard of and listened to music EVER   Own none of their
music
...                                         ...
...
188685                           Never heard of   Own none of their
music
188686                                Heard of   Own none of their
music
188687                                Heard of   Own none of their
music
188688  Heard of and listened to music RECENTLY  Own a lot of their
music
```

```
188689  Heard of and listened to music RECENTLY   Own none of their
music

        LIKE_ARTIST  words_score  GENDER   AGE  \
0               NaN         -2.0  Female  41.0
1               NaN          5.0  Female  45.0
2               NaN          1.0    Male  23.0
3              28.0          4.0  Female  61.0
4              18.0          2.0  Female  20.0
...             ...          ...     ...   ...
188685          NaN          3.0  Female  53.0
188686          NaN         -1.0    Male  52.0
188687          NaN          3.0  Female  28.0
188688         74.0         10.0  Female  35.0
188689          7.0          1.0  Female  82.0

                                  WORKING    REGION  \
0                   Temporarily unemployed     North
1                                      NaN    Centre
2             Employed 8-29 hours per week  Midlands
3             Retired from self-employment  Midlands
4                   Temporarily unemployed     South
...                                    ...       ...
188685                                 NaN     North
188686            Employed 30+ hours a week  Midlands
188687  Full-time housewife / househusband     North
188688            Employed 30+ hours a week     North
188689                                 NaN    Centre

                                                      MUSIC
LIST_OWN  \
0        Music means a lot to me and is a passion of mine            3
hours
1       Music is important to me but not necessarily m...
1
2        Music means a lot to me and is a passion of mine            5
hours
3       Music is important to me but not necessarily m...
1 hour
4       Music is important to me but not necessarily m...  Less than
an hour
...                                                   ...
...
188685  Music is important to me but not necessarily m...
1
188686  I like music but it does not feature heavily i...
1 hour
188687  Music is important to me but not necessarily m...
NaN
188688  Music is important to me but not necessarily m...
```

```
1 hour
188689  Music is important to me but not necessarily m...
0

       LIST_BACK     Q1    Q2    Q3    Q4    Q5    Q6    Q7    Q8
Q9   Q10  \
0         0 Hours   62.0  22.0  62.0  48.0  35.0  30.0  48.0  28.0
88.0  70.0
1               2   32.0  57.0  52.0  10.0  10.0  29.0  73.0  51.0
12.0  50.0
2             NaN  100.0  75.0  90.0  48.0  25.0  34.0  46.0  29.0
29.0  71.0
3             NaN   62.0  57.0  55.0  44.0  53.0  66.0  33.0  27.0
41.0  52.0
4         3 hours   22.0  69.0  28.0  52.0  32.0  22.0   9.0  10.0
11.0  55.0
...           ...    ...   ...   ...   ...   ...   ...   ...   ...   ..
.   ...
188685        NaN   68.0  52.0  66.0  49.0  49.0  31.0  30.0   8.0
29.0  49.0
188686     1 hour   75.0  50.0  32.0   7.0  48.0  50.0  66.0  30.0
48.0  48.0
188687        NaN   52.0  67.0  51.0  52.0  53.0  35.0  15.0  14.0
53.0  51.0
188688     1 hour    9.0  27.0  13.0  13.0   6.0  58.0  45.0  30.0
61.0  13.0
188689          0   73.0  92.0  93.0   7.0  11.0   9.0  11.0   9.0
34.0  73.0

          Q11   Q12    Q13    Q14   Q15   Q16   Q17   Q18   Q19
0        49.0  49.0   32.0   32.0  50.0  31.0  31.0  10.0   9.0
1        91.0  72.0   32.0   55.0  53.0  54.0  75.0   NaN   NaN
2        72.0  48.0  100.0  100.0  28.0  65.0  72.0  73.0  83.0
3        71.0  73.0   53.0   61.0  49.0  52.0  63.0  50.0  45.0
4        84.0  70.0   20.0   19.0  11.0  47.0  71.0  37.0  26.0
...       ...   ...    ...    ...   ...   ...   ...   ...   ...
188685   74.0  69.0   50.0   30.0  11.0  51.0  51.0   NaN   NaN
188686   48.0  30.0   30.0   49.0  32.0  32.0  47.0  31.0   8.0
188687   50.0  51.0   57.0   51.0  52.0  52.0  52.0  54.0  47.0
188688   54.0  49.0   65.0   50.0   4.0  46.0  77.0  47.0  39.0
188689   73.0  73.0   54.0   69.0   8.0  10.0  70.0   NaN   NaN

[188690 rows x 35 columns]
```

```
plt.title('To what extent do you like or dislike listening this
artist? vs. Rating')
sns.scatterplot(x='LIKE_ARTIST', y='Rating', hue='GENDER',
data=training_merge_df.sample(15000));
```




```
training_merge_df[training_merge_df['LIKE_ARTIST'].isna()].Rating.desc
ribe()
```

```
count    133662.000000
mean         32.326353
std          20.782582
min           0.000000
25%          12.000000
50%          30.000000
75%          48.000000
max         100.000000
Name: Rating, dtype: float64
```

```
training_merge_df.Rating.describe()
```

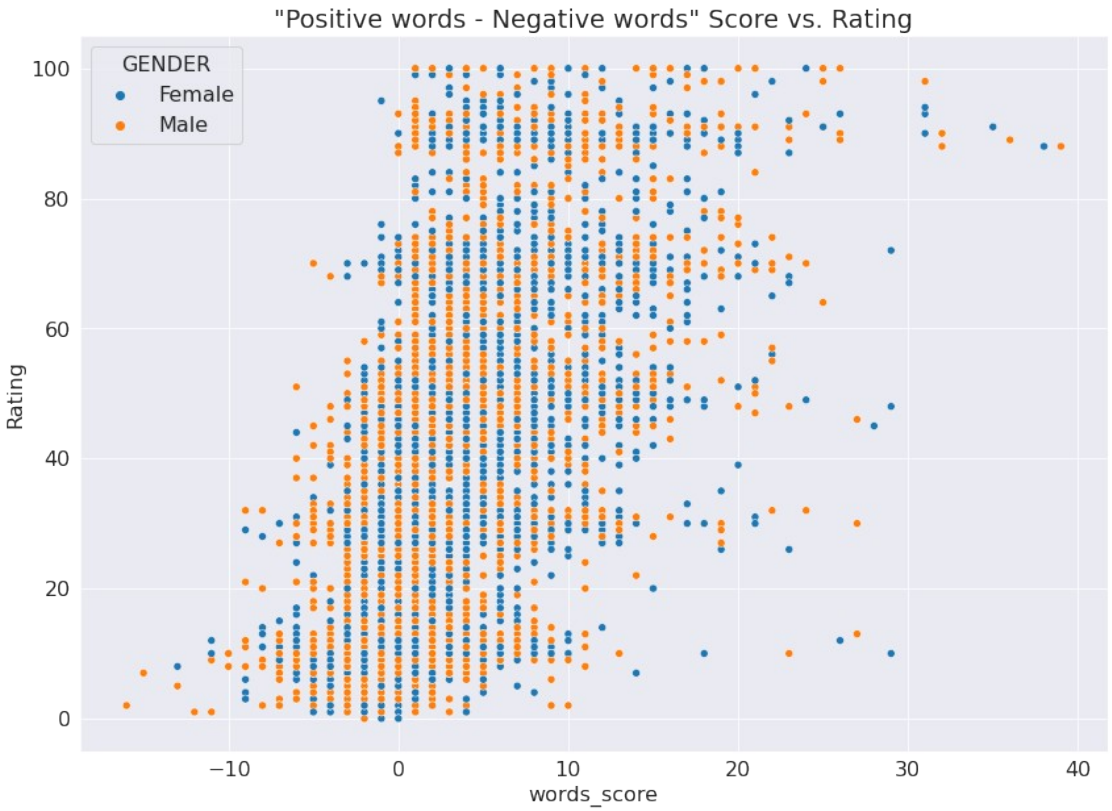
```
count    188690.000000
mean         36.435391
std          22.586036
min           0.000000
25%          15.000000
50%          32.000000
75%          50.000000
max         100.000000
Name: Rating, dtype: float64
```

```
training_merge_df[~training_merge_df['LIKE_ARTIST'].isna()].Rating.des
cribe()
```

```
count    55028.000000
mean        46.416170
std         23.653523
min          0.000000
25%         30.000000
50%         48.000000
75%         64.250000
max        100.000000
Name: Rating, dtype: float64
```

Words_Score

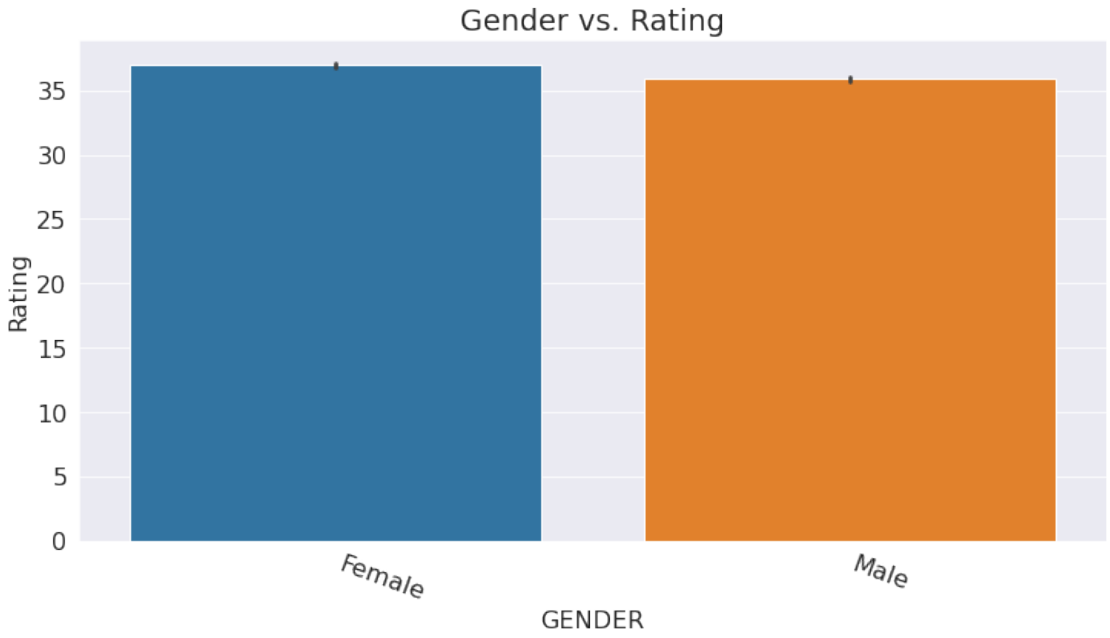
```
plt.title('"Positive words - Negative words" Score vs. Rating')
sns.scatterplot(x='words_score', y='Rating', hue='GENDER',
data=training_merge_df.sample(10000));
```



GENDER

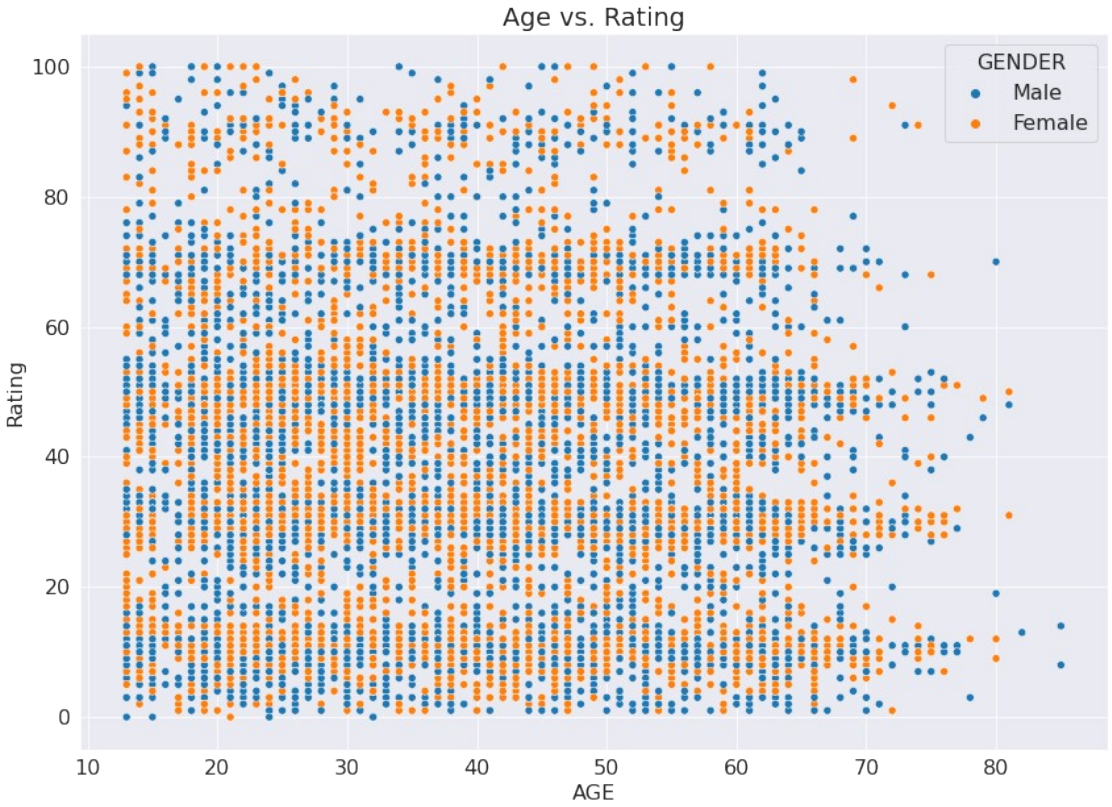
```
fig, ax = plt.subplots(figsize=(12,6))

plt.title('Gender vs. Rating')
sns.barplot(x='GENDER', y='Rating', data=training_merge_df)
```

```
plt.xticks(rotation=340, ha='left')
plt.show();
```



AGE

```
plt.title('Age vs. Rating')
sns.scatterplot(x='AGE', y='Rating', hue='GENDER',
data=training_merge_df.sample(10000));
```



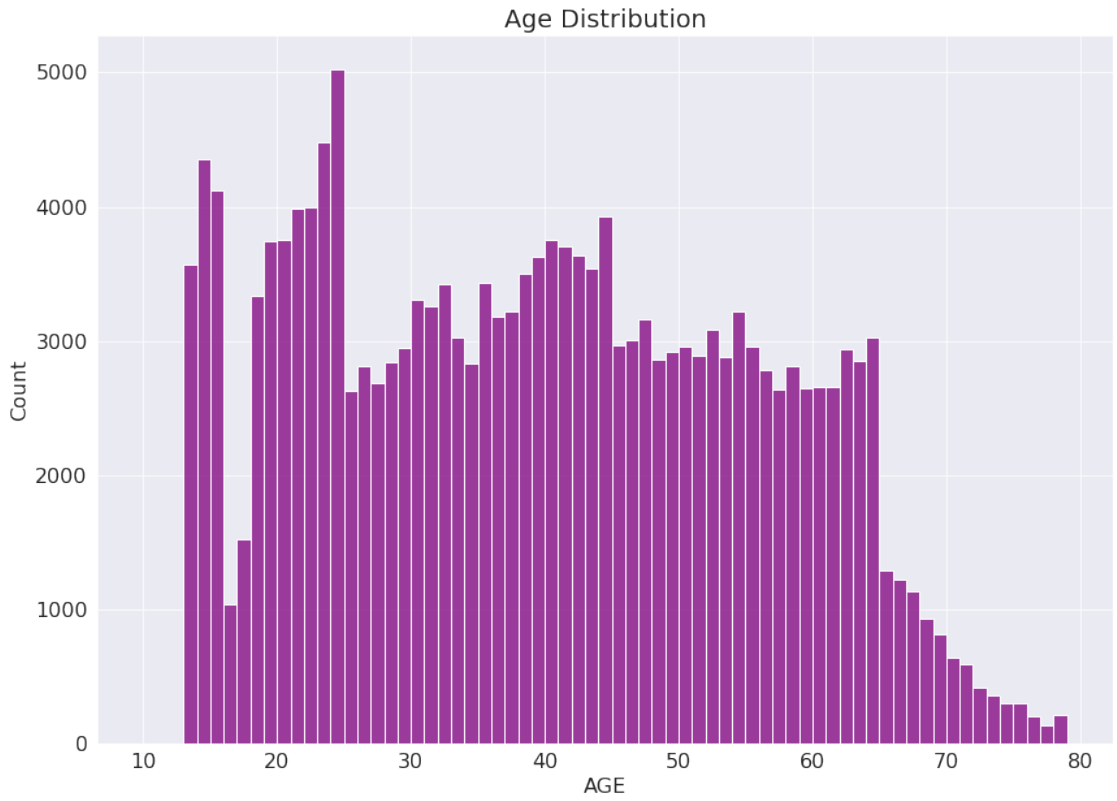

```
training_merge_df['AGE'].describe()
```

```
count    174982.000000
mean         39.246923
std          16.035515
min          13.000000
25%          25.000000
50%          39.000000
75%          52.000000
max          94.000000
Name: AGE, dtype: float64
```

```
print('Nan cells in the training_merge_df table
{}'.format(training_merge_df['AGE'].isna().sum()))
```

```
Nan cells in the training_merge_df table 13708
```

```
plt.title('Age Distribution')
sns.histplot(training_merge_df.AGE, bins=np.arange(10,80,1),
color='purple');
```




```
training_merge_df[training_merge_df['AGE'] > 50].AGE.count()
```

```
48897
```

```
def age_to_categorical(x):
  try:
    if int(x) <= 17:
      return '13-17'
    elif 17< int(x) <= 25:
      return '18-25'
    elif 25< int(x) <= 35:
      return '26-35'
    elif 35< int(x) <= 50:
      return '36-50'
    elif 50< int(x) <= 65:
      return '51-65'
    else:
      return 'older than 65'
  except:
    return np.nan

training_merge_df['AGE_GROUP'] = training_merge_df['AGE'].apply(lambda
x: age_to_categorical(x))

training_merge_df['AGE_GROUP'].value_counts()
```

```
36-50            49963
51-65            41315
18-25            30944
26-35            30573
13-17            14605
older than 65     7582
Name: AGE_GROUP, dtype: int64
```

```
training_merge_df['AGE_GROUP'].fillna('36-50', inplace=True)
training_merge_df['AGE'].fillna(39, inplace=True)
```

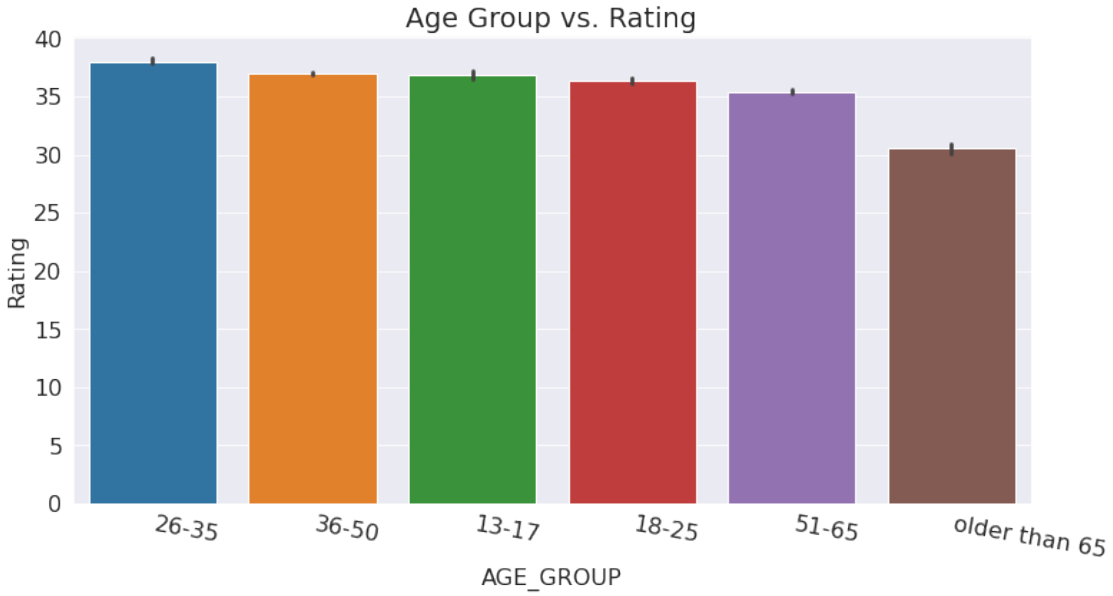
Test DataFrame

```
test_merge_df['AGE_GROUP'] = test_merge_df['AGE'].apply(lambda x:
age_to_categorical(x))
test_merge_df['AGE_GROUP'].fillna('36-50', inplace=True)
test_merge_df['AGE'].fillna(39, inplace=True)
```

```
plot_order= training_merge_df.groupby('AGE_GROUP')
['Rating'].mean().sort_values(ascending=False).index.values
```

```
fig, ax = plt.subplots(figsize=(12,6))
```

```
plt.title('Age Group vs. Rating')
sns.barplot(x='AGE_GROUP', y='Rating', data=training_merge_df,
order=plot_order)
plt.xticks(rotation=350, ha='left')
plt.show();
```

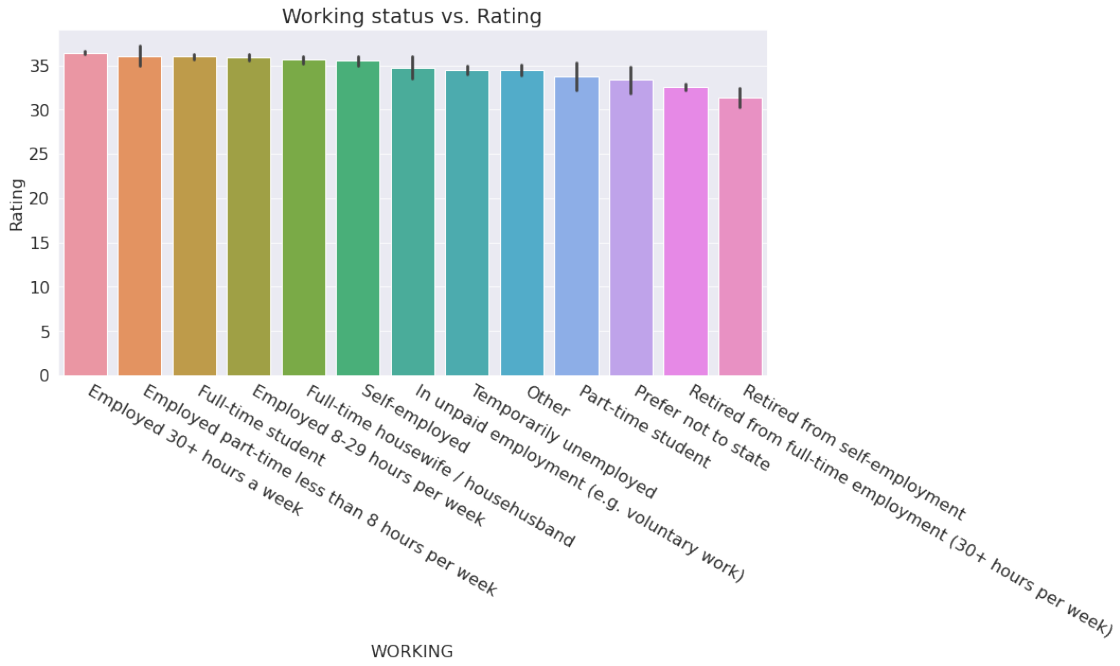


Working

```
training_merge_df['WORKING'].value_counts()
```

```
Employed 30+ hours a week                                    53347
Full-time student                                            20244
Employed 8-29 hours per week                                 16284
Retired from full-time employment (30+ hours per week)       13234
Full-time housewife / househusband                           10367
Self-employed                                                 7629
Temporarily unemployed                                        7528
Other                                                         5725
Retired from self-employment                                  1480
Employed part-time less than 8 hours per week                 1480
In unpaid employment (e.g. voluntary work)                    1407
Prefer not to state                                            947
Part-time student                                              873
Name: WORKING, dtype: int64
```

```
plot_order= training_merge_df.groupby('WORKING')
['Rating'].mean().sort_values(ascending=False).index.values

fig, ax = plt.subplots(figsize=(12,6))

plt.title('Working status vs. Rating')
sns.barplot(x='WORKING', y='Rating', data=training_merge_df,
order=plot_order)
plt.xticks(rotation=330, ha='left')
plt.show();
```



Region

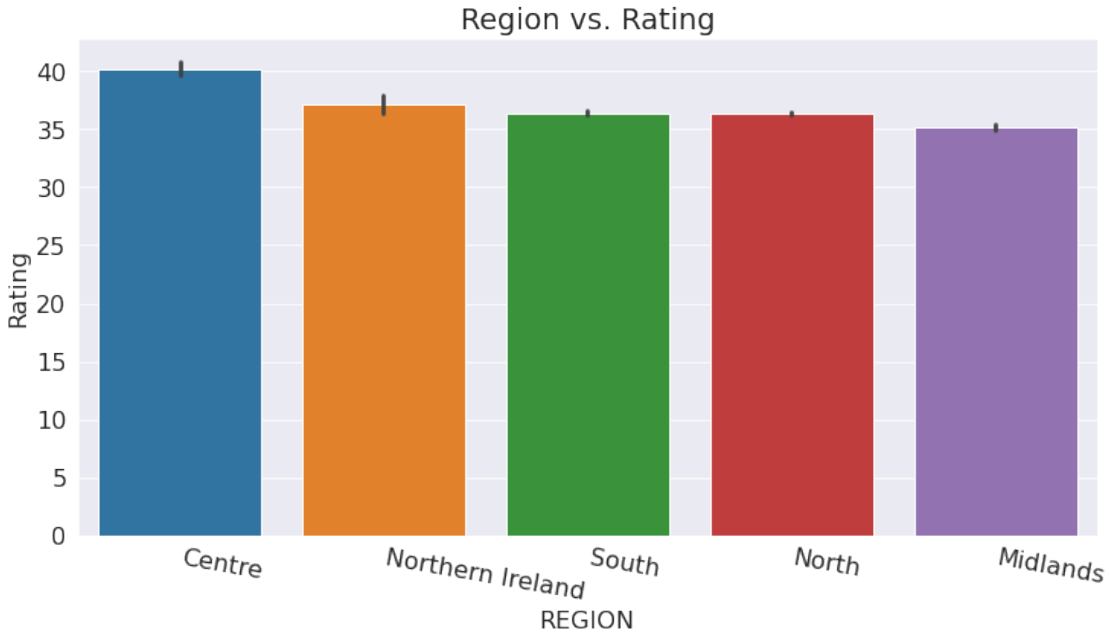
```
training_merge_df['REGION'].unique()
```

```
array(['North', 'Centre', 'Midlands', 'South', nan, 'Northern
Ireland',
       'North Ireland'], dtype=object)
```

```
training_merge_df['REGION'].value_counts()
```

```
North               58707
South               54005
Midlands            44220
Centre               7284
Northern Ireland     2890
North Ireland         375
Name: REGION, dtype: int64
```

```
training_merge_df['REGION'].replace(['North Ireland'], 'Northern
Ireland', inplace=True)
test_merge_df['REGION'].replace(['North Ireland'], 'Northern Ireland',
inplace=True)

plot_order= training_merge_df.groupby('REGION')
['Rating'].mean().sort_values(ascending=False).index.values

fig, ax = plt.subplots(figsize=(12,6))

plt.title('Region vs. Rating')
sns.barplot(x='REGION', y='Rating', data=training_merge_df,
order=plot_order)
plt.xticks(rotation=350, ha='left')
plt.show();
```



Music

```
training_merge_df['MUSIC'].unique()
```

```
array(['Music means a lot to me and is a passion of mine',
       'Music is important to me but not necessarily more important',
       'I like music but it does not feature heavily in my life',
       'Music is important to me but not necessarily more important
than other hobbies or interests',
       nan, 'Music has no particular interest for me',
       'Music is no longer as important as it used to be to me'],
      dtype=object)
```

```
training_merge_df['MUSIC'].value_counts()
```

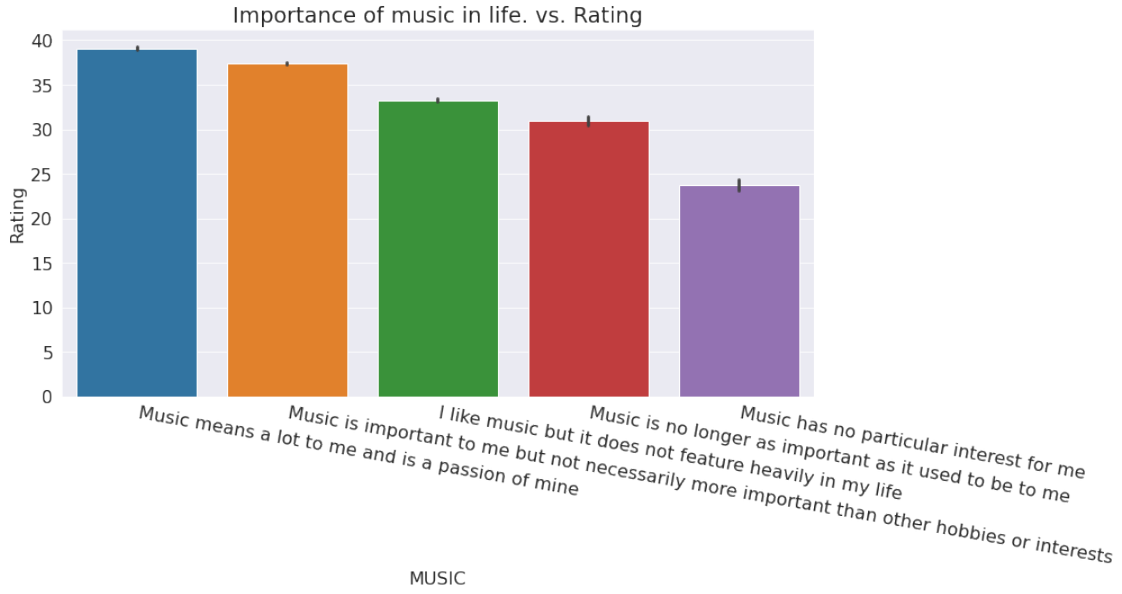
```
Music is important to me but not necessarily more important
56695
Music means a lot to me and is a passion of mine
54793
I like music but it does not feature heavily in my life
43023
Music is important to me but not necessarily more important than other
hobbies or interests    12977
Music is no longer as important as it used to be to me
5702
Music has no particular interest for me
3643
Name: MUSIC, dtype: int64
```

```
training_merge_df['MUSIC'].replace(['Music is important to me but not
necessarily more important'], 'Music is important to me but not
necessarily more important than other hobbies or interests',
inplace=True)
test_merge_df['MUSIC'].replace(['Music is important to me but not
necessarily more important'], 'Music is important to me but not
necessarily more important than other hobbies or interests',
inplace=True)
```

```
training_merge_df['MUSIC'].value_counts()
```

```
Music is important to me but not necessarily more important than other
hobbies or interests    69672
Music means a lot to me and is a passion of mine
54793
I like music but it does not feature heavily in my life
43023
Music is no longer as important as it used to be to me
5702
Music has no particular interest for me
3643
Name: MUSIC, dtype: int64
```

```
plot_order= training_merge_df.groupby('MUSIC')
['Rating'].mean().sort_values(ascending=False).index.values

fig, ax = plt.subplots(figsize=(12,6))

plt.title('Importance of music in life. vs. Rating')
sns.barplot(x='MUSIC', y='Rating', data=training_merge_df,
order=plot_order)
plt.xticks(rotation=350, ha='left')
plt.show();
```



List own

```
training_merge_df['LIST_OWN'].unique()

array(['3 hours', '1', '5 hours', '1 hour', 'Less than an hour',
       '0 Hours', nan, '2', '2 hours', '4 hours', '10 hours', '16+
hours',
       '0', '6 hours', '8 hours', '4', '3', '14 hours', '15 hours',
       '7 hours', '13 hours', '12 hours', '5', '6', '8', '10', '12',
       '9 hours', '7', '11 hours', '16 hours', '15', 'More than 16
hours',
       '20', '16', '9', '17', '14', '11', '18', '22', '24', '13'],
      dtype=object)

training_merge_df['LIST_OWN'].value_counts()

1 hour                29683
2 hours               27505
Less than an hour     26697
3 hours               13078
0 Hours               12367
```

```
1                       8801
4 hours                 8116
2                       6937
5 hours                 4430
3                       2959
0                       2792
6 hours                 2744
16+ hours               1978
8 hours                 1874
10 hours                1807
4                       1465
7 hours                 1164
5                        940
12 hours                 774
9 hours                  471
6                        361
11 hours                 235
8                        234
15 hours                 231
10                       217
14 hours                 193
16 hours                 130
7                        121
13 hours                 106
12                        94
9                         40
15                        22
14                        20
16                        17
20                        13
More than 16 hours        13
17                         7
11                         6
22                         3
13                         3
24                         2
18                         1
Name: LIST_OWN, dtype: int64
```

```
training_merge_df['LIST_OWN'].isna().sum()
```

```
30039
```

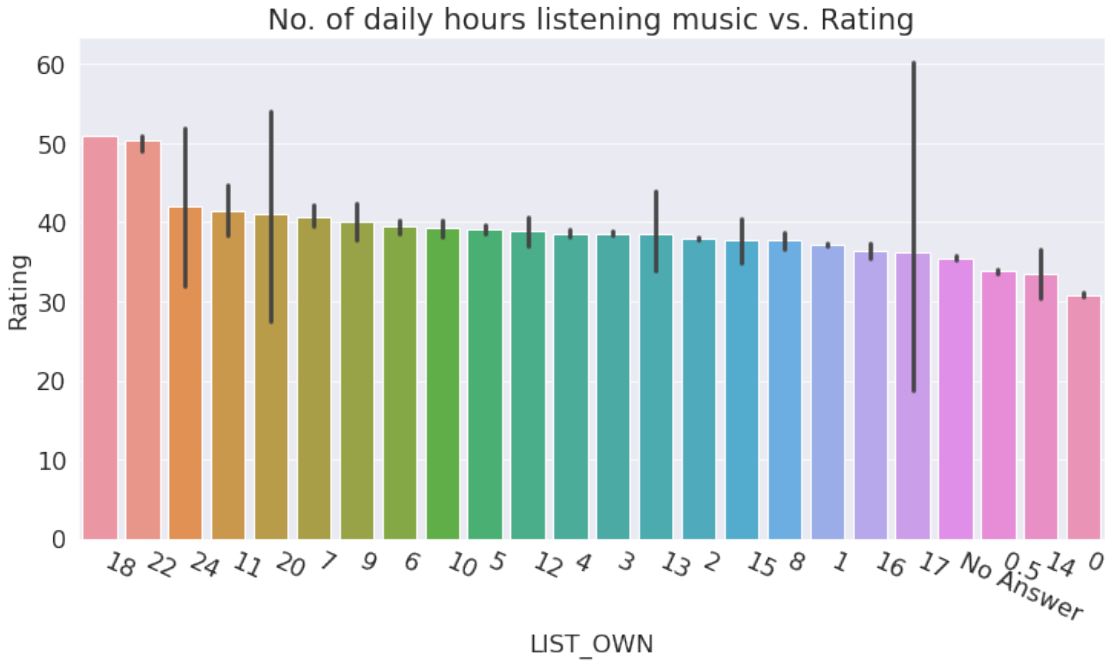
```
training_merge_df['LIST_OWN'].replace(['0 Hours'], '0', inplace=True)
training_merge_df['LIST_OWN'].replace(['Less than an hour'], '0.5',
inplace=True)
training_merge_df['LIST_OWN'].replace(['1 hour'], '1', inplace=True)
training_merge_df['LIST_OWN'].replace(['2 hours'], '2', inplace=True)
training_merge_df['LIST_OWN'].replace(['3 hours'], '3', inplace=True)
training_merge_df['LIST_OWN'].replace(['4 hours'], '4', inplace=True)
training_merge_df['LIST_OWN'].replace(['5 hours'], '5', inplace=True)
```

```
training_merge_df['LIST_OWN'].replace(['6 hours'], '6', inplace=True)
training_merge_df['LIST_OWN'].replace(['7 hours'], '7', inplace=True)
training_merge_df['LIST_OWN'].replace(['8 hours'], '8', inplace=True)
training_merge_df['LIST_OWN'].replace(['9 hours'], '9', inplace=True)
training_merge_df['LIST_OWN'].replace(['10 hours'], '10',
inplace=True)
training_merge_df['LIST_OWN'].replace(['11 hours'], '11',
inplace=True)
training_merge_df['LIST_OWN'].replace(['12 hours'], '12',
inplace=True)
training_merge_df['LIST_OWN'].replace(['13 hours'], '13',
inplace=True)
training_merge_df['LIST_OWN'].replace(['14 hours'], '14',
inplace=True)
training_merge_df['LIST_OWN'].replace(['15 hours'], '15',
inplace=True)
training_merge_df['LIST_OWN'].replace(['16 hours'], '16',
inplace=True)
training_merge_df['LIST_OWN'].replace(['16+ hours'], '16',
inplace=True)
training_merge_df['LIST_OWN'].replace(['More than 16 hours'], '16',
inplace=True)

training_merge_df['LIST_OWN'].fillna('No Answer', inplace=True)
```

Test DataFrame

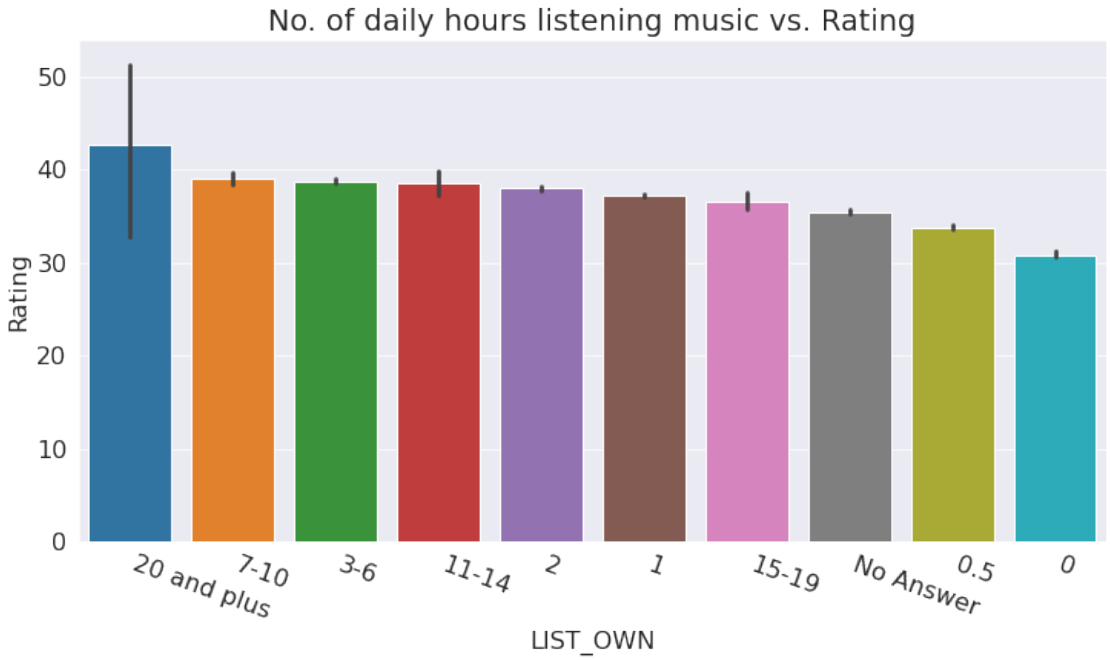
```
test_merge_df['LIST_OWN'].replace(['0 Hours'], '0', inplace=True)
test_merge_df['LIST_OWN'].replace(['Less than an hour'], '0.5',
inplace=True)
test_merge_df['LIST_OWN'].replace(['1 hour'], '1', inplace=True)
test_merge_df['LIST_OWN'].replace(['2 hours'], '2', inplace=True)
test_merge_df['LIST_OWN'].replace(['3 hours'], '3', inplace=True)
test_merge_df['LIST_OWN'].replace(['4 hours'], '4', inplace=True)
test_merge_df['LIST_OWN'].replace(['5 hours'], '5', inplace=True)
test_merge_df['LIST_OWN'].replace(['6 hours'], '6', inplace=True)
test_merge_df['LIST_OWN'].replace(['7 hours'], '7', inplace=True)
test_merge_df['LIST_OWN'].replace(['8 hours'], '8', inplace=True)
test_merge_df['LIST_OWN'].replace(['9 hours'], '9', inplace=True)
test_merge_df['LIST_OWN'].replace(['10 hours'], '10', inplace=True)
test_merge_df['LIST_OWN'].replace(['11 hours'], '11', inplace=True)
test_merge_df['LIST_OWN'].replace(['12 hours'], '12', inplace=True)
test_merge_df['LIST_OWN'].replace(['13 hours'], '13', inplace=True)
test_merge_df['LIST_OWN'].replace(['14 hours'], '14', inplace=True)
test_merge_df['LIST_OWN'].replace(['15 hours'], '15', inplace=True)
test_merge_df['LIST_OWN'].replace(['16 hours'], '16', inplace=True)
test_merge_df['LIST_OWN'].replace(['16+ hours'], '16', inplace=True)
test_merge_df['LIST_OWN'].replace(['More than 16 hours'], '16',
inplace=True)
test_merge_df['LIST_OWN'].fillna('No Answer', inplace=True)
```

```
plot_order= training_merge_df.groupby('LIST_OWN')
['Rating'].mean().sort_values(ascending=False).index.values

fig, ax = plt.subplots(figsize=(12,6))

plt.title('No. of daily hours listening music vs. Rating')
sns.barplot(x='LIST_OWN', y='Rating', data=training_merge_df,
order=plot_order)
plt.xticks(rotation=335, ha='left')
plt.show();
```




```
lo_mapper = {'No Answer': 'No Answer',
             '0': '0',
             '0.5': '0.5',
             '1': '1',
             '2': '2',
             '3': '3-6',
             '4': '3-6',
             '5': '3-6',
             '6': '3-6',
             '7': '7-10',
             '8': '7-10',
             '9': '7-10',
             '10': '7-10',
             '11': '11-14',
             '12':  '11-14',
             '13': '11-14',
             '14': '11-14',
             '15': '15-19',
             '16': '15-19',
```

```
            '17': '15-19',
            '18': '15-19',
            '19': '15-19',
            '20': '20 and plus',
            '21': '20 and plus',
            '22': '20 and plus',
            '23': '20 and plus',
            '24': '20 and plus'
            }

training_merge_df['LIST_OWN'] =
training_merge_df['LIST_OWN'].map(lo_mapper)

training_merge_df['LIST_OWN'].unique()

array(['3-6', '1', '0.5', '0', 'No Answer', '2', '7-10', '15-19', '11-
14',
       '20 and plus'], dtype=object)

training_merge_df['LIST_OWN'].value_counts()

1              38484
2              34442
3-6            34093
No Answer      30039
0.5            26697
0              15159
7-10            5928
15-19           2399
11-14           1431
20 and plus       18
Name: LIST_OWN, dtype: int64

plot_order= training_merge_df.groupby('LIST_OWN')
['Rating'].mean().sort_values(ascending=False).index.values

fig, ax = plt.subplots(figsize=(12,6))

plt.title('No. of daily hours listening music vs. Rating')
sns.barplot(x='LIST_OWN', y='Rating', data=training_merge_df,
order=plot_order)
plt.xticks(rotation=340, ha='left')
plt.show();
```




```
test_merge_df['LIST_OWN'] = test_merge_df['LIST_OWN'].map(lo_mapper)
```

List Back

```
training_merge_df['LIST_BACK'].unique()
```

```
array(['0 Hours', '2', nan, '3 hours', 'Less than an hour', '4 hours',
       '8 hours', '5 hours', '4', '3', '1 hour', '2 hours', '5', '1',
       '6 hours', '7 hours', 'More than 16 hours', '0', '9 hours',
'6',
       '14 hours', '16+ hours', '8', '10 hours', '9', '16 hours',
       '15 hours', '12', '12 hours', '10', '20', '18', '11 hours',
       '13 hours', '7', '14', '15', '19', '24', '16', '11', '21'],
      dtype=object)
```

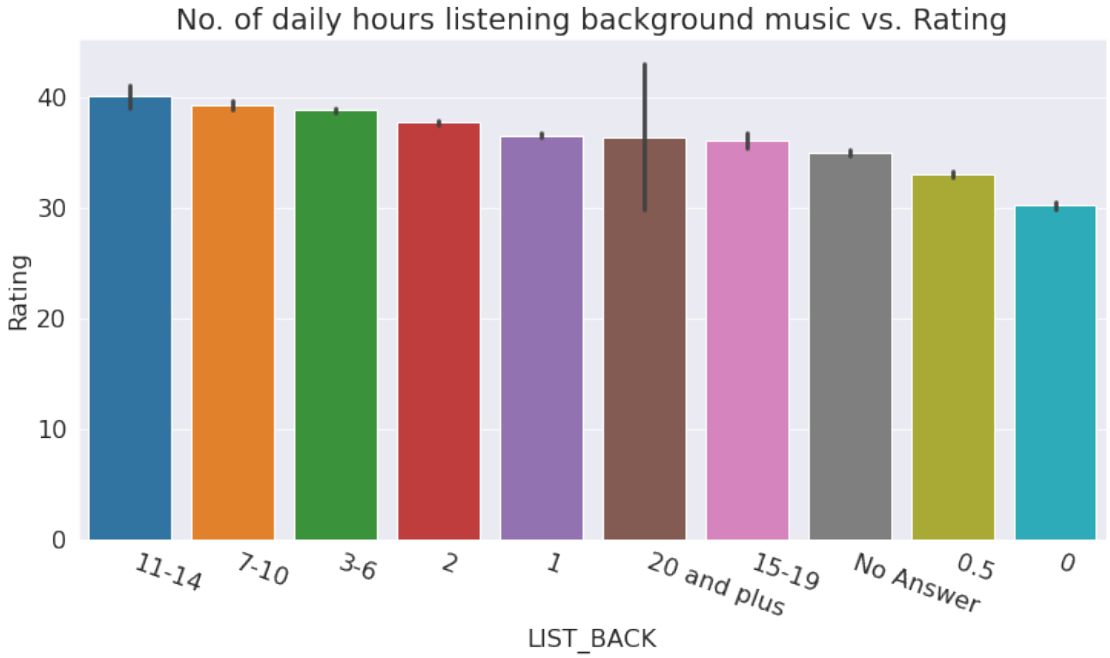
```
training_merge_df['LIST_BACK'].value_counts()
```

```
2 hours               24663
1 hour                23409
Less than an hour     22232
3 hours               13679
0 Hours               10565
4 hours               10492
1                      6856
5 hours                6170
2                      6027
6 hours                5099
8 hours                4473
3                      3097
16+ hours              2890
```

```
0                       2768
7 hours                 2572
10 hours                2544
4                       2284
5                       1587
9 hours                 1200
12 hours                1171
6                       1119
8                       1013
7                        498
14 hours                 369
11 hours                 334
15 hours                 325
10                       319
16 hours                 278
12                       213
13 hours                 213
9                        189
20                        36
More than 16 hours        23
15                        17
14                        14
19                        11
24                        11
16                        11
11                        10
18                         8
21                         1
Name: LIST_BACK, dtype: int64
```

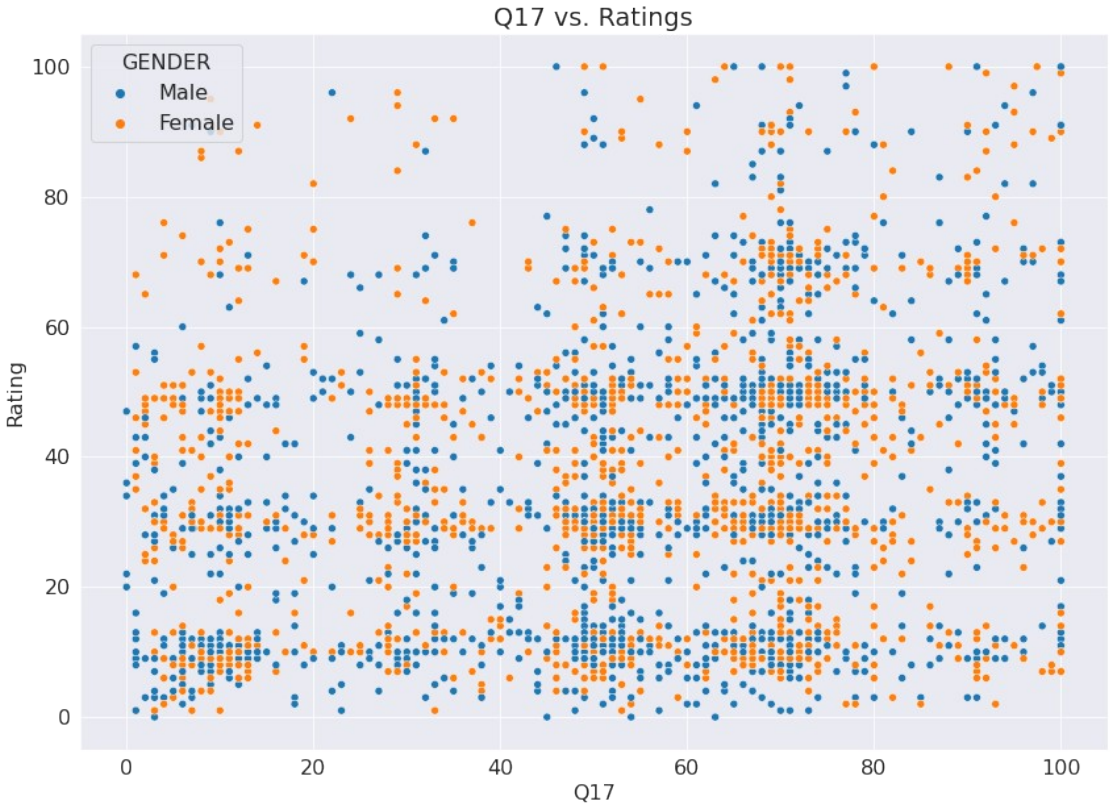
```
training_merge_df['LIST_BACK'].replace(['0 Hours'], '0', inplace=True)
training_merge_df['LIST_BACK'].replace(['Less than an hour'], '0.5',
inplace=True)
training_merge_df['LIST_BACK'].replace(['1 hour'], '1', inplace=True)
training_merge_df['LIST_BACK'].replace(['2 hours'], '2', inplace=True)
training_merge_df['LIST_BACK'].replace(['3 hours'], '3', inplace=True)
training_merge_df['LIST_BACK'].replace(['4 hours'], '4', inplace=True)
training_merge_df['LIST_BACK'].replace(['5 hours'], '5', inplace=True)
training_merge_df['LIST_BACK'].replace(['6 hours'], '6', inplace=True)
training_merge_df['LIST_BACK'].replace(['7 hours'], '7', inplace=True)
training_merge_df['LIST_BACK'].replace(['8 hours'], '8', inplace=True)
training_merge_df['LIST_BACK'].replace(['9 hours'], '9', inplace=True)
training_merge_df['LIST_BACK'].replace(['10 hours'], '10',
inplace=True)
training_merge_df['LIST_BACK'].replace(['11 hours'], '11',
inplace=True)
training_merge_df['LIST_BACK'].replace(['12 hours'], '12',
inplace=True)
training_merge_df['LIST_BACK'].replace(['13 hours'], '13',
inplace=True)
```

```
training_merge_df['LIST_BACK'].replace(['14 hours'], '14',
inplace=True)
training_merge_df['LIST_BACK'].replace(['15 hours'], '15',
inplace=True)
training_merge_df['LIST_BACK'].replace(['16 hours'], '16',
inplace=True)
training_merge_df['LIST_BACK'].replace(['16+ hours'], '16',
inplace=True)
training_merge_df['LIST_BACK'].replace(['More than 16 hours'], '16',
inplace=True)

training_merge_df['LIST_BACK'].fillna('No Answer', inplace=True)

training_merge_df['LIST_BACK'] =
training_merge_df['LIST_BACK'].map(lo_mapper)

plot_order= training_merge_df.groupby('LIST_BACK')
['Rating'].mean().sort_values(ascending=False).index.values

fig, ax = plt.subplots(figsize=(12,6))

plt.title('No. of daily hours listening background music vs. Rating')
sns.barplot(x='LIST_BACK', y='Rating', data=training_merge_df,
order=plot_order)
plt.xticks(rotation=340, ha='left')
plt.show();
```



Test DataFrame

```
test_merge_df['LIST_BACK'].replace(['0 Hours'], '0', inplace=True)
test_merge_df['LIST_BACK'].replace(['Less than an hour'], '0.5',
inplace=True)
test_merge_df['LIST_BACK'].replace(['1 hour'], '1', inplace=True)
test_merge_df['LIST_BACK'].replace(['2 hours'], '2', inplace=True)
test_merge_df['LIST_BACK'].replace(['3 hours'], '3', inplace=True)
test_merge_df['LIST_BACK'].replace(['4 hours'], '4', inplace=True)
test_merge_df['LIST_BACK'].replace(['5 hours'], '5', inplace=True)
test_merge_df['LIST_BACK'].replace(['6 hours'], '6', inplace=True)
test_merge_df['LIST_BACK'].replace(['7 hours'], '7', inplace=True)
test_merge_df['LIST_BACK'].replace(['8 hours'], '8', inplace=True)
test_merge_df['LIST_BACK'].replace(['9 hours'], '9', inplace=True)
test_merge_df['LIST_BACK'].replace(['10 hours'], '10', inplace=True)
test_merge_df['LIST_BACK'].replace(['11 hours'], '11', inplace=True)
test_merge_df['LIST_BACK'].replace(['12 hours'], '12', inplace=True)
test_merge_df['LIST_BACK'].replace(['13 hours'], '13', inplace=True)
test_merge_df['LIST_BACK'].replace(['14 hours'], '14', inplace=True)
test_merge_df['LIST_BACK'].replace(['15 hours'], '15', inplace=True)
test_merge_df['LIST_BACK'].replace(['16 hours'], '16', inplace=True)
test_merge_df['LIST_BACK'].replace(['16+ hours'], '16', inplace=True)
test_merge_df['LIST_BACK'].replace(['More than 16 hours'], '16',
inplace=True)
test_merge_df['LIST_BACK'].fillna('No Answer', inplace=True)

test_merge_df['LIST_BACK'] = test_merge_df['LIST_BACK'].map(lo_mapper)

plt.title('Q17 vs. Ratings')
sns.scatterplot(x='Q17', y='Rating', hue='GENDER',
data=training_merge_df.sample(3000));
```




```
training_merge_df.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 188690 entries, 0 to 188689
Data columns (total 36 columns):
 #   Column             Non-Null Count   Dtype
---  ------             --------------   -----
 0   Artist             188690 non-null  int64
 1   Track              188690 non-null  int64
 2   User               188690 non-null  int64
 3   Rating             188690 non-null  int64
 4   Time               188690 non-null  int64
 5   HEARD_OF           188690 non-null  object
 6   OWN_ARTIST_MUSIC   188690 non-null  object
 7   LIKE_ARTIST        55028 non-null   float64
 8   words_score        186636 non-null  float64
 9   GENDER             176833 non-null  object
 10  AGE                188690 non-null  float64
 11  WORKING            140545 non-null  object
 12  REGION             167481 non-null  object
 13  MUSIC              176833 non-null  object
 14  LIST_OWN           188690 non-null  object
 15  LIST_BACK          188690 non-null  object
 16  Q1                 176833 non-null  float64
 17  Q2                 176833 non-null  float64
 18  Q3                 176833 non-null  float64
```

```
 19  Q4                176833 non-null  float64
 20  Q5                176833 non-null  float64
 21  Q6                176833 non-null  float64
 22  Q7                176833 non-null  float64
 23  Q8                176833 non-null  float64
 24  Q9                176833 non-null  float64
 25  Q10               176833 non-null  float64
 26  Q11               176833 non-null  float64
 27  Q12               176833 non-null  float64
 28  Q13               176833 non-null  float64
 29  Q14               176833 non-null  float64
 30  Q15               176833 non-null  float64
 31  Q16               142754 non-null  float64
 32  Q17               176833 non-null  float64
 33  Q18               140545 non-null  float64
 34  Q19               140545 non-null  float64
 35  AGE_GROUP         188690 non-null  object
dtypes: float64(22), int64(5), object(9)
memory usage: 57.3+ MB
```

HEARD_OF

```
mapper = {'Heard of and listened to music RECENTLY': 4,
          'Heard of and listened to music EVER': 3,
          'Heard of': 2,
          'Never heard of': 1}

training_merge_df['HEARD_OF'] =
training_merge_df['HEARD_OF'].map(mapper)
```

Test DataFrame

```
test_merge_df['HEARD_OF'] = test_merge_df['HEARD_OF'].map(mapper)

training_merge_df['HEARD_OF'].unique()

array([1, 3, 2, 4])

training_merge_df['HEARD_OF'].value_counts()

1    98169
2    35493
3    34990
4    20038
Name: HEARD_OF, dtype: int64
```

Own Art Music

```
oam_mapper = {'Own all or most of their music': 4,
          'Own a lot of their music': 3,
```

```
          'Own a little of their music': 2,
          'Own none of their music': 1}

training_merge_df['OWN_ARTIST_MUSIC'] =
training_merge_df['OWN_ARTIST_MUSIC'].map(oam_mapper)
```

Test DataFrame

```
test_merge_df['OWN_ARTIST_MUSIC'] =
test_merge_df['OWN_ARTIST_MUSIC'].map(oam_mapper)

training_merge_df['OWN_ARTIST_MUSIC'].unique()
```

```
array([1, 2, 4, 3])
```

```
training_merge_df['OWN_ARTIST_MUSIC'].value_counts()
```

```
1    160113
2     18721
3      7263
4      2593
Name: OWN_ARTIST_MUSIC, dtype: int64
```

Like Artist

```
training_merge_df['LIKE_ARTIST'].isna().sum()
```

```
133662
```

```
def to_categorical(x):
  try:
    if 1<= int(x) <= 10:
      return '1-10'
    elif 11<= int(x) <= 20:
      return '11-20'
    elif 21<= int(x) <= 30:
      return '21-30'
    elif 31<= int(x) <= 40:
      return '31-40'
    elif 41<= int(x) <= 50:
      return '41-50'
    elif 51<= int(x) <= 60:
      return '51-60'
    elif 61<= int(x) <= 70:
      return '61-70'
    elif 71<= int(x) <= 80:
      return '71-80'
    elif 81<= int(x) <= 90:
      return '81-90'
    else:
      return '91-100'
```

```
  except:
    return np.nan

training_merge_df['LIKE_ARTIST'] =
training_merge_df['LIKE_ARTIST'].apply(lambda x: to_categorical(x))

test_merge_df['LIKE_ARTIST'] =
test_merge_df['LIKE_ARTIST'].apply(lambda x: to_categorical(x))

training_merge_df['LIKE_ARTIST'].fillna('No Answer', inplace=True)

test_merge_df['LIKE_ARTIST'].fillna('No Answer', inplace=True)

training_merge_df['LIKE_ARTIST'].value_counts()
```

```
No Answer    133662
41-50         11114
21-30          8804
51-60          8244
31-40          6574
61-70          6415
71-80          4976
1-10           2825
91-100         2269
11-20          2126
81-90          1681
Name: LIKE_ARTIST, dtype: int64
```

Music

```
training_merge_df['MUSIC'].unique()
```

```
array(['Music means a lot to me and is a passion of mine',
       'Music is important to me but not necessarily more important
than other hobbies or interests',
       'I like music but it does not feature heavily in my life', nan,
       'Music has no particular interest for me',
       'Music is no longer as important as it used to be to me'],
      dtype=object)
```

```
training_merge_df['MUSIC'].value_counts()
```

```
Music is important to me but not necessarily more important than other
hobbies or interests    69672
Music means a lot to me and is a passion of mine
54793
I like music but it does not feature heavily in my life
43023
Music is no longer as important as it used to be to me
5702
Music has no particular interest for me
```

```
3643
Name: MUSIC, dtype: int64
```

```
m_mapper = {'Music means a lot to me and is a passion of mine': 6,
          'Music is important to me but not necessarily more important
than other hobbies or interests': 5,
          'No Answer': 4,
          'I like music but it does not feature heavily in my life':
3,
          'Music is no longer as important as it used to be to me': 2,
          'Music has no particular interest for me': 1,
          }

training_merge_df['MUSIC'] = training_merge_df['MUSIC'].map(m_mapper)
```

Test DataFrame

```
test_merge_df['MUSIC'] = test_merge_df['MUSIC'].map(m_mapper)
```

#Missing Values in DF

```
training_merge_df['GENDER'].fillna('No Answer', inplace=True)
training_merge_df['WORKING'].fillna('No Answer', inplace=True)
training_merge_df['REGION'].fillna('No Answer', inplace=True)

test_merge_df['GENDER'].fillna('No Answer', inplace=True)
test_merge_df['WORKING'].fillna('No Answer', inplace=True)
test_merge_df['REGION'].fillna('No Answer', inplace=True)
```

#Training & Validation Sets

As test set is already given.

We put 20% of Training test into calidation set.

```
from sklearn.model_selection import train_test_split

training_df, validation_df = train_test_split(training_merge_df,
test_size=0.2)

print('training_df.shape :', training_df.shape)
print('validation_df.shape :', validation_df.shape)
```

```
training_df.shape : (150952, 36)
validation_df.shape : (37738, 36)
```

```
training_df
```

```
        Artist  Track   User  Rating  Time  HEARD_OF  OWN_ARTIST_MUSIC
\
168265      35     88  30594      69    23         1                 1

186415      15     41  16939      30     9         2                 1
```

```
186064      48    172  47900      28    17         1                  1

38552       26     63  23658      79    22         1                  1

149111      23     57  21142      28    21         1                  1

...        ...    ...    ...     ...   ...       ...                ...

31991       45    163  45329      31    16         3                  1

25672       16    134  34029      13    12         1                  1

81988        6     14   5979      68     7         1                  1

97594       15     33  13060      14    19         1                  1

53332       28     72  23225      31    22         1                  1


       LIKE_ARTIST  words_score     GENDER   AGE  \
168265   No Answer         -1.0     Female  37.0
186415   No Answer          2.0       Male  21.0
186064   No Answer          3.0       Male  34.0
38552    No Answer         11.0  No Answer  39.0
149111   No Answer         -3.0     Female  47.0
...            ...          ...        ...   ...
31991        21-30          0.0       Male  63.0
25672    No Answer          0.0     Female  63.0
81988    No Answer          3.0       Male  34.0
97594    No Answer         -2.0       Male  53.0
53332    No Answer         -2.0  No Answer  39.0

                                                  WORKING     REGION
MUSIC  \
168265                Full-time housewife / househusband   Midlands
3.0
186415                                     Full-time student      South
6.0
186064                          Employed 30+ hours a week      South
5.0
38552                                          No Answer  No Answer
NaN
149111                       Employed 8-29 hours per week   Midlands
6.0
...                                                 ...        ...
...
31991                                     Self-employed      North
5.0
```

```
25672   Retired from full-time employment (30+ hours p...      North
3.0
81988                                              No Answer      North
6.0
97594                              Employed 30+ hours a week      North
3.0
53332                                              No Answer  No Answer
NaN

         LIST_OWN  LIST_BACK    Q1     Q2    Q3    Q4    Q5    Q6
Q7    Q8  \
168265          1          1  53.0  100.0  35.0  12.0  12.0  36.0
14.0  14.0
186415          2        3-6  79.0   73.0  84.0  57.0  35.0  16.0
28.0  21.0
186064        0.5          0  35.0   47.0  35.0  48.0  33.0   3.0
35.0  29.0
38552   No Answer  No Answer   NaN    NaN   NaN   NaN   NaN   NaN
NaN   NaN
149111          2        3-6  58.0   62.0  55.0  30.0  29.0  51.0
25.0  20.0
...           ...        ...   ...    ...   ...   ...   ...   ...   ..
.   ...
31991           0          0  50.0   50.0  11.0  11.0  51.0  94.0
9.0   9.0
25672         0.5       7-10  10.0   49.0  48.0  48.0  48.0  46.0
71.0  72.0
81988         3-6        3-6  47.0   46.0  48.0  47.0  48.0  48.0
48.0  49.0
97594           1        0.5  14.0   47.0  12.0  34.0  69.0   8.0
6.0   5.0
53332   No Answer  No Answer   NaN    NaN   NaN   NaN   NaN   NaN
NaN   NaN

          Q9   Q10   Q11   Q12    Q13   Q14   Q15   Q16    Q17   Q18
Q19  \
168265  50.0  68.0  65.0  51.0  100.0  86.0   4.0   NaN  100.0   4.0
52.0
186415  44.0  67.0  73.0  62.0   88.0  67.0  48.0  70.0   69.0  60.0
30.0
186064  72.0  60.0  51.0  53.0   41.0  45.0   6.0   3.0   48.0  34.0
33.0
38552    NaN   NaN   NaN   NaN    NaN   NaN   NaN   NaN    NaN   NaN
NaN
149111  51.0  99.0  67.0  69.0   46.0  46.0  43.0  20.0   86.0  22.0
34.0
...      ...   ...   ...   ...    ...   ...   ...   ...    ...   ...
...
31991   70.0  10.0  34.0  10.0   10.0  11.0  11.0  13.0   55.0  10.0
11.0
```

```
25672   71.0  46.0  48.0  51.0   51.0  31.0  96.0   7.0   51.0   8.0
9.0
81988   55.0  54.0  55.0  54.0   53.0  54.0  53.0  55.0   55.0   NaN
NaN
97594   75.0  52.0  57.0  21.0   15.0  27.0  11.0  14.0   75.0  18.0
13.0
53332    NaN   NaN   NaN   NaN    NaN   NaN   NaN   NaN    NaN   NaN
NaN

       AGE_GROUP
168265     36-50
186415     18-25
186064     26-35
38552      36-50
149111     36-50
...          ...
31991      51-65
25672      51-65
81988      26-35
97594      51-65
53332      36-50

[150952 rows x 36 columns]

validation_df

        Artist  Track   User  Rating  Time  HEARD_OF  OWN_ARTIST_MUSIC
\
109566       4     11   5357      74    18         2                 1

49742       20     44  17177      53    21         1                 1

92733       28     73  23226      49    22         3                 2

38465       22    122  32594      54     0         4                 2

20298       31     79  26606      12    11         1                 1

...        ...    ...    ...     ...   ...       ...               ...

84579       26     64  22187      31    22         1                 1

47851       10    145  35978       9    12         3                 1

165143      37     97  30961      46    23         4                 2

51210       46    168  44138      11    16         3                 1

116670      46    165  44448      30    16         1                 1
```

```
       LIKE_ARTIST  words_score     GENDER   AGE  \
109566   No Answer          7.0       Male  55.0
49742    No Answer          7.0       Male  15.0
92733        41-50          2.0  No Answer  39.0
38465       91-100         21.0       Male  54.0
20298    No Answer         -3.0       Male  56.0
...            ...          ...        ...   ...
84579    No Answer          5.0     Female  39.0
47851        11-20         -1.0       Male  22.0
165143       71-80          5.0     Female  14.0
51210        21-30          0.0       Male  41.0
116670   No Answer          3.0     Female  42.0

                              WORKING     REGION  MUSIC   LIST_OWN
LIST_BACK  \
109566     Employed 30+ hours a week   Midlands    5.0          1
3-6
49742               Full-time student      South    5.0          1
2
92733                      No Answer  No Answer    NaN  No Answer  No
Answer
38465                      No Answer  No Answer    5.0          1
7-10
20298      Employed 30+ hours a week      South    5.0          1
3-6
...                              ...        ...    ...        ...
...
84579      Employed 30+ hours a week      South    6.0        3-6
7-10
47851               Full-time student      North    5.0          1
0.5
165143                 Self-employed      South    6.0        3-6
3-6
51210      Employed 30+ hours a week   Midlands    6.0        3-6
7-10
116670  Employed 8-29 hours per week      South    3.0        0.5
3-6

           Q1    Q2    Q3    Q4    Q5    Q6    Q7    Q8    Q9    Q10
Q11  \
109566   31.0  86.0  54.0  26.0   4.0   5.0   4.0   3.0  55.0   88.0
89.0
49742    29.0  28.0  17.0  44.0  45.0  65.0  43.0  31.0  31.0   32.0
64.0
92733     NaN   NaN   NaN   NaN   NaN   NaN   NaN   NaN   NaN    NaN
NaN
38465    53.0  54.0  55.0  55.0  69.0  37.0  67.0  65.0  56.0   32.0
57.0
```

```
20298    15.0  14.0  14.0  11.0  13.0  13.0  14.0  12.0  51.0   51.0
92.0
...       ...   ...   ...   ...   ...   ...   ...   ...   ...    ...
...
84579    79.0  83.0  79.0  35.0   6.0   3.0   5.0  55.0   6.0  100.0
57.0
47851   100.0  49.0  49.0  26.0  11.0  25.0  11.0  11.0  29.0   47.0
100.0
165143   85.0  76.0  74.0  87.0  66.0  48.0  57.0  44.0  22.0   66.0
72.0
51210    52.0  64.0  46.0  69.0  40.0  95.0  19.0  21.0  20.0   67.0
49.0
116670   30.0  52.0  30.0  30.0  29.0  31.0  53.0  28.0  29.0   30.0
70.0

          Q12    Q13    Q14   Q15    Q16   Q17   Q18   Q19 AGE_GROUP
109566   63.0    4.0   51.0  51.0  100.0  55.0  72.0   7.0     51-65
49742    61.0   44.0   61.0  22.0   29.0  49.0  50.0  27.0     13-17
92733     NaN    NaN    NaN   NaN    NaN   NaN   NaN   NaN     36-50
38465    56.0   55.0   54.0  69.0   32.0  54.0   NaN   NaN     51-65
20298    92.0   53.0   53.0  10.0   12.0  14.0  12.0  12.0     51-65
...       ...    ...    ...   ...    ...   ...   ...   ...       ...
84579    54.0   62.0   59.0  56.0    NaN  72.0  65.0  56.0     36-50
47851   100.0  100.0  100.0   9.0   30.0  52.0  29.0  29.0     18-25
165143   64.0   52.0   88.0  85.0    NaN  79.0  77.0  76.0     13-17
51210    66.0   72.0   71.0  73.0    4.0  78.0  53.0  52.0     36-50
116670   69.0   31.0   51.0  30.0   30.0  70.0  30.0  30.0     36-50

[37738 rows x 36 columns]
```

Input and Target Col's

```
input_cols = list(training_df.columns)
input_cols.remove('Rating')
input_cols.remove('AGE')

target_col = 'Rating'

training_inputs = training_df[input_cols].copy()
training_targets = training_df[target_col].copy()

validation_inputs = validation_df[input_cols].copy()
validation_targets = validation_df[target_col].copy()

test_inputs = test_merge_df[input_cols].copy()

training_inputs
```

```
        Artist  Track   User  Time  HEARD_OF  OWN_ARTIST_MUSIC
LIKE_ARTIST  \
168265      35     88  30594    23         1                 1   No
```

```
Answer
186415      15     41  16939     9         2                  1   No
Answer
186064      48    172  47900    17         1                  1   No
Answer
38552       26     63  23658    22         1                  1   No
Answer
149111      23     57  21142    21         1                  1   No
Answer
...        ...    ...    ...   ...       ...                ...
...
31991       45    163  45329    16         3                  1
21-30
25672       16    134  34029    12         1                  1   No
Answer
81988        6     14   5979     7         1                  1   No
Answer
97594       15     33  13060    19         1                  1   No
Answer
53332       28     72  23225    22         1                  1   No
Answer

        words_score     GENDER  \
168265         -1.0     Female
186415          2.0       Male
186064          3.0       Male
38552          11.0  No Answer
149111         -3.0     Female
...             ...        ...
31991           0.0       Male
25672           0.0     Female
81988           3.0       Male
97594          -2.0       Male
53332          -2.0  No Answer

                                                    WORKING     REGION
MUSIC  \
168265                 Full-time housewife / househusband   Midlands
3.0
186415                                  Full-time student      South
6.0
186064                        Employed 30+ hours a week      South
5.0
38552                                       No Answer  No Answer
NaN
149111                      Employed 8-29 hours per week   Midlands
6.0
...                                                ...        ...
...
31991                                  Self-employed      North
```

```
5.0
25672   Retired from full-time employment (30+ hours p...      North
3.0
81988                                              No Answer      North
6.0
97594                              Employed 30+ hours a week      North
3.0
53332                                              No Answer  No Answer
NaN

         LIST_OWN  LIST_BACK    Q1     Q2    Q3    Q4    Q5    Q6
Q7    Q8  \
168265          1          1  53.0  100.0  35.0  12.0  12.0  36.0
14.0  14.0
186415          2        3-6  79.0   73.0  84.0  57.0  35.0  16.0
28.0  21.0
186064        0.5          0  35.0   47.0  35.0  48.0  33.0   3.0
35.0  29.0
38552   No Answer  No Answer   NaN    NaN   NaN   NaN   NaN   NaN
NaN   NaN
149111          2        3-6  58.0   62.0  55.0  30.0  29.0  51.0
25.0  20.0
...           ...        ...   ...    ...   ...   ...   ...   ...   ..
.   ...
31991           0          0  50.0   50.0  11.0  11.0  51.0  94.0
9.0   9.0
25672         0.5       7-10  10.0   49.0  48.0  48.0  48.0  46.0
71.0  72.0
81988         3-6        3-6  47.0   46.0  48.0  47.0  48.0  48.0
48.0  49.0
97594           1        0.5  14.0   47.0  12.0  34.0  69.0   8.0
6.0   5.0
53332   No Answer  No Answer   NaN    NaN   NaN   NaN   NaN   NaN
NaN   NaN

          Q9   Q10   Q11   Q12    Q13   Q14   Q15   Q16    Q17   Q18
Q19  \
168265  50.0  68.0  65.0  51.0  100.0  86.0   4.0   NaN  100.0   4.0
52.0
186415  44.0  67.0  73.0  62.0   88.0  67.0  48.0  70.0   69.0  60.0
30.0
186064  72.0  60.0  51.0  53.0   41.0  45.0   6.0   3.0   48.0  34.0
33.0
38552    NaN   NaN   NaN   NaN    NaN   NaN   NaN   NaN    NaN   NaN
NaN
149111  51.0  99.0  67.0  69.0   46.0  46.0  43.0  20.0   86.0  22.0
34.0
...      ...   ...   ...   ...    ...   ...   ...   ...    ...   ...
...
31991   70.0  10.0  34.0  10.0   10.0  11.0  11.0  13.0   55.0  10.0
```

```
11.0
25672   71.0  46.0  48.0  51.0   51.0  31.0  96.0   7.0   51.0   8.0
9.0
81988   55.0  54.0  55.0  54.0   53.0  54.0  53.0  55.0   55.0   NaN
NaN
97594   75.0  52.0  57.0  21.0   15.0  27.0  11.0  14.0   75.0  18.0
13.0
53332    NaN   NaN   NaN   NaN    NaN   NaN   NaN   NaN    NaN   NaN
NaN

       AGE_GROUP
168265     36-50
186415     18-25
186064     26-35
38552      36-50
149111     36-50
...          ...
31991      51-65
25672      51-65
81988      26-35
97594      51-65
53332      36-50

[150952 rows x 34 columns]

validation_inputs

        Artist  Track   User  Time  HEARD_OF  OWN_ARTIST_MUSIC
LIKE_ARTIST  \
109566       4     11   5357    18         2                 1   No
Answer
49742       20     44  17177    21         1                 1   No
Answer
92733       28     73  23226    22         3                 2
41-50
38465       22    122  32594     0         4                 2
91-100
20298       31     79  26606    11         1                 1   No
Answer
...        ...    ...    ...   ...       ...               ...
...
84579       26     64  22187    22         1                 1   No
Answer
47851       10    145  35978    12         3                 1
11-20
165143      37     97  30961    23         4                 2
71-80
51210       46    168  44138    16         3                 1
21-30
116670      46    165  44448    16         1                 1   No
```

Answer

	words_score	GENDER	WORKING	REGION
109566	7.0	Male	Employed 30+ hours a week	Midlands
49742	7.0	Male	Full-time student	South
92733	2.0	No Answer	No Answer	No Answer
38465	21.0	Male	No Answer	No Answer
20298	-3.0	Male	Employed 30+ hours a week	South
...	...	...	...	...
84579	5.0	Female	Employed 30+ hours a week	South
47851	-1.0	Male	Full-time student	North
165143	5.0	Female	Self-employed	South
51210	0.0	Male	Employed 30+ hours a week	Midlands
116670	3.0	Female	Employed 8-29 hours per week	South

	MUSIC	LIST_OWN	LIST_BACK	Q1	Q2	Q3	Q4	Q5	Q6
109566	5.0	1	3-6	31.0	86.0	54.0	26.0	4.0	5.0
49742	5.0	1	2	29.0	28.0	17.0	44.0	45.0	65.0
92733	NaN	No Answer	No Answer	NaN	NaN	NaN	NaN	NaN	NaN
38465	5.0	1	7-10	53.0	54.0	55.0	55.0	69.0	37.0
20298	5.0	1	3-6	15.0	14.0	14.0	11.0	13.0	13.0
...	...	...	...	...	...	...	...	...	...
84579	6.0	3-6	7-10	79.0	83.0	79.0	35.0	6.0	3.0
47851	5.0	1	0.5	100.0	49.0	49.0	26.0	11.0	25.0
165143	6.0	3-6	3-6	85.0	76.0	74.0	87.0	66.0	48.0
51210	6.0	3-6	7-10	52.0	64.0	46.0	69.0	40.0	95.0
116670	3.0	0.5	3-6	30.0	52.0	30.0	30.0	29.0	

```
31.0

          Q7    Q8    Q9    Q10    Q11    Q12    Q13    Q14   Q15
Q16  \
109566   4.0   3.0  55.0   88.0   89.0   63.0    4.0   51.0  51.0
100.0
49742   43.0  31.0  31.0   32.0   64.0   61.0   44.0   61.0  22.0
29.0
92733    NaN   NaN   NaN    NaN    NaN    NaN    NaN    NaN   NaN
NaN
38465   67.0  65.0  56.0   32.0   57.0   56.0   55.0   54.0  69.0
32.0
20298   14.0  12.0  51.0   51.0   92.0   92.0   53.0   53.0  10.0
12.0
...      ...   ...   ...    ...    ...    ...    ...    ...   ...    .
..
84579    5.0  55.0   6.0  100.0   57.0   54.0   62.0   59.0  56.0
NaN
47851   11.0  11.0  29.0   47.0  100.0  100.0  100.0  100.0   9.0
30.0
165143  57.0  44.0  22.0   66.0   72.0   64.0   52.0   88.0  85.0
NaN
51210   19.0  21.0  20.0   67.0   49.0   66.0   72.0   71.0  73.0
4.0
116670  53.0  28.0  29.0   30.0   70.0   69.0   31.0   51.0  30.0
30.0

         Q17   Q18   Q19 AGE_GROUP
109566  55.0  72.0   7.0     51-65
49742   49.0  50.0  27.0     13-17
92733    NaN   NaN   NaN     36-50
38465   54.0   NaN   NaN     51-65
20298   14.0  12.0  12.0     51-65
...      ...   ...   ...       ...
84579   72.0  65.0  56.0     36-50
47851   52.0  29.0  29.0     18-25
165143  79.0  77.0  76.0     13-17
51210   78.0  53.0  52.0     36-50
116670  70.0  30.0  30.0     36-50

[37738 rows x 34 columns]

test_inputs

        Artist  Track   User  Time  HEARD_OF  OWN_ARTIST_MUSIC
LIKE_ARTIST  \
0            1      6   3475    18         3                 1
1-10
1            6    149  39210    15         1                 1   No
Answer
```

```
2           40    177  47861    17         1                  1   No
Answer
3           31     79  27413    11         1                  1   No
Answer
4           26     66  23232    22         1                  1   No
Answer
...        ...    ...    ...   ...       ...                ...
...
125789      14     95  30004    23         2                  1   No
Answer
125790      10     25   8186     7         1                  1   No
Answer
125791      40    146  38180    13         2                  1   No
Answer
125792      22    113  32918     0         3                  1
41-50
125793       2     70  24231    22         1                  1   No
Answer

        words_score     GENDER
WORKING  \
0               2.0     Female                      Employed 30+ hours
a week
1               NaN       Male                      Employed 30+ hours
a week
2              -2.0     Female
Other
3               0.0     Female  Employed part-time less than 8 hours
per week
4               0.0  No Answer                                   No
Answer
...             ...        ...
...
125789         12.0       Male                      Employed 30+ hours
a week
125790          6.0       Male                                  No
Answer
125791          3.0     Female             Full-time housewife /
househusband
125792          2.0     Female                                  No
Answer
125793          4.0       Male                      Employed 30+ hours
a week

           REGION  MUSIC   LIST_OWN  LIST_BACK    Q1    Q2     Q3
Q4    Q5  \
0           South    6.0          1        3-6   8.0  69.0   27.0
27.0  50.0
1       Midlands    5.0          1          1  81.0  67.0   94.0
61.0  53.0
```

```
2        Midlands    2.0        0.5        0.5   9.0  94.0   49.0
48.0  49.0
3        Midlands    3.0          1          1  53.0  38.0   51.0
53.0  53.0
4       No Answer    NaN  No Answer  No Answer   NaN   NaN    NaN
NaN   NaN
...           ...    ...        ...        ...   ...   ...    ...   ..
.   ...
125789   Midlands    6.0       7-10        3-6  84.0  69.0  100.0
32.0   9.0
125790      North    3.0  No Answer        3-6  29.0  70.0   30.0
30.0  69.0
125791   Midlands    6.0      15-19       7-10  59.0  51.0   51.0
83.0  32.0
125792  No Answer    6.0          0          1  69.0  30.0   76.0
74.0  73.0
125793      North    5.0          1        3-6  15.0  68.0   51.0
51.0  51.0

          Q6    Q7    Q8    Q9   Q10   Q11   Q12   Q13   Q14   Q15
Q16  \
0       27.0  26.0   8.0  51.0  50.0  66.0  49.0  20.0   7.0   8.0
9.0
1       32.0  41.0  42.0  36.0  76.0  70.0  76.0  58.0  61.0  66.0
51.0
2        8.0  13.0  56.0  92.0  92.0  55.0  57.0  11.0  57.0  10.0
11.0
3       53.0  33.0  51.0  47.0  33.0  41.0  45.0  49.0  49.0  49.0
49.0
4        NaN   NaN   NaN   NaN   NaN   NaN   NaN   NaN   NaN   NaN
NaN
...      ...   ...   ...   ...   ...   ...   ...   ...   ...   ...   .
..
125789  28.0   9.0  12.0  50.0  75.0  68.0  72.0  64.0  70.0  75.0
NaN
125790  14.0  12.0  12.0  70.0  29.0  50.0  48.0  54.0  66.0  10.0
34.0
125791  43.0  14.0  41.0  71.0  58.0  36.0  43.0  81.0  63.0  45.0
65.0
125792  11.0  11.0  11.0  92.0  34.0  74.0  72.0  36.0  37.0   9.0
9.0
125793  71.0   2.0   2.0  94.0  65.0   2.0   3.0   3.0   3.0   3.0
NaN

          Q17   Q18   Q19 AGE_GROUP
0        7.0   4.0   8.0     36-50
1       75.0  70.0  72.0     26-35
2       91.0   7.0   9.0     51-65
3       35.0  52.0  52.0     18-25
4        NaN   NaN   NaN     36-50
```

```
...      ...   ...   ...       ...
125789  72.0  56.0  54.0     36-50
125790  70.0   NaN   NaN     36-50
125791  30.0  46.0  21.0     36-50
125792  64.0   NaN   NaN     36-50
125793  30.0   5.0   5.0     36-50

[125794 rows x 34 columns]
```

Segregation of Numeric and Catego... Cols

```
numeric_cols = ['Artist', 'Track', 'User', 'Time', 'HEARD_OF',
'OWN_ARTIST_MUSIC', 'words_score', 'MUSIC', 'Q1',
'Q2','Q3','Q4','Q5','Q6','Q7','Q8','Q9','Q10','Q11','Q12','Q13','Q14',
'Q15','Q16','Q17','Q18','Q19']

categorical_cols = ['LIKE_ARTIST', 'GENDER', 'WORKING', 'REGION',
'LIST_OWN', 'LIST_BACK', 'AGE_GROUP' ]

training_inputs[numeric_cols].describe()
```

```
              Artist          Track           User           Time  \
count  150952.000000  150952.000000  150952.000000  150952.000000
mean       22.206688      86.473912   26463.279082      15.656050
std        14.478913      55.988137   13628.240967       6.443697
min         0.000000       0.000000       0.000000       0.000000
25%        10.000000      36.000000   17700.000000      12.000000
50%        22.000000      80.000000   27805.000000      17.000000
75%        35.000000     142.000000   35924.000000      21.000000
max        49.000000     183.000000   50927.000000      23.000000

            HEARD_OF  OWN_ARTIST_MUSIC    words_score
MUSIC  \
count  150952.000000     150952.000000  149299.000000  141389.000000

mean        1.877252          1.217314       2.684285       4.643445

std         1.057053          0.573414       4.841826       1.333523

min         1.000000          1.000000     -16.000000       1.000000

25%         1.000000          1.000000       0.000000       3.000000

50%         1.000000          1.000000       2.000000       5.000000

75%         3.000000          1.000000       5.000000       6.000000

max         4.000000          4.000000      39.000000       6.000000
```

	Q1	Q2	Q3	Q4	\
count	141389.000000	141389.000000	141389.000000	141389.000000	
mean	49.046179	54.551589	51.256483	37.331147	
std	27.626250	23.826195	26.497574	23.634200	
min	0.000000	0.000000	0.000000	0.000000	
25%	28.000000	43.000000	31.000000	14.000000	
50%	51.000000	53.000000	52.000000	34.000000	
75%	70.000000	71.000000	71.000000	52.000000	
max	100.000000	100.000000	100.000000	100.000000	

	Q5	Q6	Q7	Q8	\
count	141389.000000	141389.000000	141389.000000	141389.000000	
mean	34.558534	39.268697	33.975129	29.185177	
std	23.263877	25.748706	25.777516	24.241491	
min	0.000000	0.000000	0.000000	0.000000	
25%	12.000000	14.000000	10.000000	9.000000	
50%	32.000000	35.000000	30.000000	23.000000	
75%	51.000000	53.000000	52.000000	49.000000	
max	100.000000	100.000000	100.000000	100.000000	

	Q9	Q10	Q11	Q12	\
count	141389.000000	141389.000000	141389.000000	141389.000000	
mean	47.797154	54.877063	58.602724	53.569845	
std	27.354361	25.452177	23.878980	25.393843	
min	0.000000	0.000000	0.000000	0.000000	
25%	28.000000	40.000000	48.000000	35.000000	
50%	50.000000	53.000000	64.000000	53.000000	
75%	70.000000	72.000000	73.000000	71.000000	
max	100.000000	100.000000	100.000000	100.000000	

	Q13	Q14	Q15	Q16	\
count	141389.000000	141389.000000	141389.000000	114178.000000	
mean	47.040583	53.379278	39.532544	35.832825	
std	26.763479	25.876483	26.019576	25.427310	
min	0.000000	0.000000	0.000000	0.000000	
25%	28.000000	33.000000	13.000000	11.000000	
50%	50.000000	53.000000	36.000000	32.000000	
75%	68.000000	71.000000	55.000000	52.000000	
max	100.000000	100.000000	100.000000	100.000000	

	Q17	Q18	Q19
count	141389.000000	112310.000000	112310.000000
mean	53.798933	42.284300	41.325660
std	25.913654	25.698482	26.473832
min	0.000000	0.000000	0.000000
25%	35.000000	17.000000	14.000000
50%	56.000000	47.000000	45.000000
75%	71.000000	58.000000	57.000000
max	100.000000	100.000000	100.000000

```
training_inputs[numeric_cols].info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 150952 entries, 168265 to 53332
Data columns (total 27 columns):
 #   Column            Non-Null Count   Dtype  
---  ------            --------------   -----  
 0   Artist            150952 non-null  int64  
 1   Track             150952 non-null  int64  
 2   User              150952 non-null  int64  
 3   Time              150952 non-null  int64  
 4   HEARD_OF          150952 non-null  int64  
 5   OWN_ARTIST_MUSIC  150952 non-null  int64  
 6   words_score       149299 non-null  float64
 7   MUSIC             141389 non-null  float64
 8   Q1                141389 non-null  float64
 9   Q2                141389 non-null  float64
 10  Q3                141389 non-null  float64
 11  Q4                141389 non-null  float64
 12  Q5                141389 non-null  float64
 13  Q6                141389 non-null  float64
 14  Q7                141389 non-null  float64
 15  Q8                141389 non-null  float64
 16  Q9                141389 non-null  float64
 17  Q10               141389 non-null  float64
 18  Q11               141389 non-null  float64
 19  Q12               141389 non-null  float64
 20  Q13               141389 non-null  float64
 21  Q14               141389 non-null  float64
 22  Q15               141389 non-null  float64
 23  Q16               114178 non-null  float64
 24  Q17               141389 non-null  float64
 25  Q18               112310 non-null  float64
 26  Q19               112310 non-null  float64
dtypes: float64(21), int64(6)
memory usage: 32.2 MB
```

```
training_inputs[categorical_cols].nunique()
```

```
LIKE_ARTIST    11
GENDER          3
WORKING        14
REGION          6
LIST_OWN       10
LIST_BACK      10
AGE_GROUP       6
dtype: int64
```

Replacing Missing Data

```
training_merge_df[numeric_cols].isna().sum()
```

```
Artist                  0
Track                   0
User                    0
Time                    0
HEARD_OF                0
OWN_ARTIST_MUSIC        0
words_score          2054
MUSIC               11857
Q1                  11857
Q2                  11857
Q3                  11857
Q4                  11857
Q5                  11857
Q6                  11857
Q7                  11857
Q8                  11857
Q9                  11857
Q10                 11857
Q11                 11857
Q12                 11857
Q13                 11857
Q14                 11857
Q15                 11857
Q16                 45936
Q17                 11857
Q18                 48145
Q19                 48145
dtype: int64
```

```
from sklearn.impute import SimpleImputer

imputer = SimpleImputer(strategy='mean')

imputer.fit(training_merge_df[numeric_cols])

SimpleImputer()

training_inputs[numeric_cols] =
imputer.transform(training_inputs[numeric_cols])
validation_inputs[numeric_cols] =
imputer.transform(validation_inputs[numeric_cols])
test_inputs[numeric_cols] =
imputer.transform(test_inputs[numeric_cols])

training_inputs[numeric_cols].isna().sum()

Artist              0
Track               0
```

```
User                0
Time                0
HEARD_OF            0
OWN_ARTIST_MUSIC    0
words_score         0
MUSIC               0
Q1                  0
Q2                  0
Q3                  0
Q4                  0
Q5                  0
Q6                  0
Q7                  0
Q8                  0
Q9                  0
Q10                 0
Q11                 0
Q12                 0
Q13                 0
Q14                 0
Q15                 0
Q16                 0
Q17                 0
Q18                 0
Q19                 0
dtype: int64
```

Scaling of Numeric Col's

```
from sklearn.preprocessing import MinMaxScaler

scaler = MinMaxScaler()

scaler.fit(training_merge_df[numeric_cols])

MinMaxScaler()

training_inputs[numeric_cols] =
scaler.transform(training_inputs[numeric_cols])
validation_inputs[numeric_cols] =
scaler.transform(validation_inputs[numeric_cols])
test_inputs[numeric_cols] =
scaler.transform(test_inputs[numeric_cols])

training_inputs[numeric_cols].describe()
```

```
              Artist          Track           User           Time  \
count  150952.000000  150952.000000  150952.000000  150952.000000
mean        0.453198       0.472535       0.519632       0.680698
std         0.295488       0.305946       0.267603       0.280161
min         0.000000       0.000000       0.000000       0.000000
```

```
25%         0.204082       0.196721       0.347556       0.521739
50%         0.448980       0.437158       0.545978       0.739130
75%         0.714286       0.775956       0.705402       0.913043
max         1.000000       1.000000       1.000000       1.000000

            HEARD_OF  OWN_ARTIST_MUSIC    words_score
MUSIC  \
count  150952.000000     150952.000000  150952.000000  150952.000000

mean        0.292417          0.072438       0.339714       0.728698

std         0.352351          0.191138       0.087550       0.258118

min         0.000000          0.000000       0.000000       0.000000

25%         0.000000          0.000000       0.290909       0.400000

50%         0.000000          0.000000       0.327273       0.800000

75%         0.666667          0.000000       0.381818       1.000000

max         1.000000          1.000000       1.000000       1.000000


                  Q1             Q2             Q3             Q4  \
count  150952.000000  150952.000000  150952.000000  150952.000000
mean        0.490447       0.545506       0.512547       0.373323
std         0.267368       0.230591       0.256445       0.228733
min         0.000000       0.000000       0.000000       0.000000
25%         0.290000       0.460000       0.320000       0.150000
50%         0.500000       0.545362       0.512284       0.360000
75%         0.700000       0.710000       0.700000       0.520000
max         1.000000       1.000000       1.000000       1.000000

                  Q5             Q6             Q7             Q8  \
count  150952.000000  150952.000000  150952.000000  150952.000000
mean        0.345585       0.392703       0.339740       0.291840
std         0.225149       0.249198       0.249476       0.234611
min         0.000000       0.000000       0.000000       0.000000
25%         0.120000       0.160000       0.100000       0.100000
50%         0.330000       0.390000       0.320000       0.280000
75%         0.510000       0.520000       0.510000       0.480000
max         1.000000       1.000000       1.000000       1.000000

                  Q9            Q10            Q11            Q12  \
count  150952.000000  150952.000000  150952.000000  150952.000000
mean        0.477973       0.548777       0.586040       0.535700
std         0.264737       0.246328       0.231102       0.245763
min         0.000000       0.000000       0.000000       0.000000
```

```
25%         0.290000       0.450000       0.490000       0.400000
50%         0.490000       0.540000       0.600000       0.535720
75%         0.690000       0.710000       0.720000       0.710000
max         1.000000       1.000000       1.000000       1.000000

                Q13            Q14            Q15            Q16  \
count  150952.000000  150952.000000  150952.000000  150952.000000
mean        0.470381       0.533772       0.395320       0.358427
std         0.259019       0.250434       0.251819       0.221142
min         0.000000       0.000000       0.000000       0.000000
25%         0.290000       0.350000       0.140000       0.150000
50%         0.490000       0.533462       0.395239       0.358732
75%         0.670000       0.710000       0.540000       0.500000
max         1.000000       1.000000       1.000000       1.000000

                Q17            Q18            Q19
count  150952.000000  150952.000000  150952.000000
mean        0.537964       0.422666       0.413127
std         0.250794       0.221665       0.228353
min         0.000000       0.000000       0.000000
25%         0.390000       0.300000       0.280000
50%         0.537583       0.422153       0.412752
75%         0.710000       0.520000       0.520000
max         1.000000       1.000000       1.000000
```

```
training_inputs[numeric_cols].info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 150952 entries, 168265 to 53332
Data columns (total 27 columns):
 #   Column            Non-Null Count   Dtype
---  ------            --------------   -----
 0   Artist            150952 non-null  float64
 1   Track             150952 non-null  float64
 2   User              150952 non-null  float64
 3   Time              150952 non-null  float64
 4   HEARD_OF          150952 non-null  float64
 5   OWN_ARTIST_MUSIC  150952 non-null  float64
 6   words_score       150952 non-null  float64
 7   MUSIC             150952 non-null  float64
 8   Q1                150952 non-null  float64
 9   Q2                150952 non-null  float64
 10  Q3                150952 non-null  float64
 11  Q4                150952 non-null  float64
 12  Q5                150952 non-null  float64
 13  Q6                150952 non-null  float64
 14  Q7                150952 non-null  float64
 15  Q8                150952 non-null  float64
 16  Q9                150952 non-null  float64
 17  Q10               150952 non-null  float64
```

```
 18  Q11               150952 non-null  float64
 19  Q12               150952 non-null  float64
 20  Q13               150952 non-null  float64
 21  Q14               150952 non-null  float64
 22  Q15               150952 non-null  float64
 23  Q16               150952 non-null  float64
 24  Q17               150952 non-null  float64
 25  Q18               150952 non-null  float64
 26  Q19               150952 non-null  float64
dtypes: float64(27)
memory usage: 32.2 MB
```

```
# Encoding Categorical data
```

```
training_merge_df[categorical_cols].isna().sum()
```

```
LIKE_ARTIST    0
GENDER         0
WORKING        0
REGION         0
LIST_OWN       0
LIST_BACK      0
AGE_GROUP      0
dtype: int64
```

```
training_merge_df[categorical_cols].nunique()
```

```
LIKE_ARTIST    11
GENDER          3
WORKING        14
REGION          6
LIST_OWN       10
LIST_BACK      10
AGE_GROUP       6
dtype: int64
```

```
from sklearn.preprocessing import OneHotEncoder
```

```
encoder = OneHotEncoder(sparse=False, handle_unknown='ignore')
```

```
encoder.fit(training_merge_df[categorical_cols])
```

```
OneHotEncoder(handle_unknown='ignore', sparse=False)
```

```
encoded_cols = list(encoder.get_feature_names(categorical_cols));
```

```
/usr/local/lib/python3.8/dist-packages/sklearn/utils/
deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is
deprecated in 1.0 and will be removed in 1.2. Please use
```

```
get_feature_names_out instead.


training_inputs[encoded_cols] =
encoder.transform(training_inputs[categorical_cols])
validation_inputs[encoded_cols] =
encoder.transform(validation_inputs[categorical_cols])
test_inputs[encoded_cols] =
encoder.transform(test_inputs[categorical_cols])

training_inputs

          Artist     Track      User      Time  HEARD_OF
OWN_ARTIST_MUSIC  \
168265  0.714286  0.480874  0.600742  1.000000  0.000000
0.0
186415  0.306122  0.224044  0.332613  0.391304  0.333333
0.0
186064  0.979592  0.939891  0.940562  0.739130  0.000000
0.0
38552   0.530612  0.344262  0.464547  0.956522  0.000000
0.0
149111  0.469388  0.311475  0.415143  0.913043  0.000000
0.0
...          ...       ...       ...       ...       ...
...
31991   0.918367  0.890710  0.890078  0.695652  0.666667
0.0
25672   0.326531  0.732240  0.668192  0.521739  0.000000
0.0
81988   0.122449  0.076503  0.117403  0.304348  0.000000
0.0
97594   0.306122  0.180328  0.256446  0.826087  0.000000
0.0
53332   0.571429  0.393443  0.456045  0.956522  0.000000
0.0

       LIKE_ARTIST  words_score     GENDER  \
168265   No Answer     0.272727     Female
186415   No Answer     0.327273       Male
186064   No Answer     0.345455       Male
38552    No Answer     0.490909  No Answer
149111   No Answer     0.236364     Female
...            ...          ...        ...
31991        21-30     0.290909       Male
25672    No Answer     0.290909     Female
81988    No Answer     0.345455       Male
97594    No Answer     0.254545       Male
53332    No Answer     0.254545  No Answer
```

		WORKING	
168265	Full-time housewife / househusband	Midlands	
186415	Full-time student	South	
186064	Employed 30+ hours a week	South	
38552	No Answer	No Answer	
149111	Employed 8-29 hours per week	Midlands	
...	...	...	
31991	Self-employed	North	
25672	Retired from full-time employment (30+ hours p...	North	
81988	No Answer	North	
97594	Employed 30+ hours a week	North	
53332	No Answer	No Answer	

	MUSIC	LIST_OWN	LIST_BACK	Q1	Q2	Q3
168265	0.400000	1	1	0.530000	1.000000	0.350000
186415	1.000000	2	3-6	0.790000	0.730000	0.840000
186064	0.800000	0.5	0	0.350000	0.470000	0.350000
38552	0.728824	No Answer	No Answer	0.490234	0.545362	0.512284
149111	1.000000	2	3-6	0.580000	0.620000	0.550000
...	...	...	...	...	...	...
31991	0.800000	0	0	0.500000	0.500000	0.110000
25672	0.400000	0.5	7-10	0.100000	0.490000	0.480000
81988	1.000000	3-6	3-6	0.470000	0.460000	0.480000
97594	0.400000	1	0.5	0.140000	0.470000	0.120000
53332	0.728824	No Answer	No Answer	0.490234	0.545362	0.512284

```
              Q4        Q5        Q6        Q7        Q8        Q9
Q10  \
168265  0.120000  0.120000  0.360000  0.140000  0.140000  0.500000
0.680000
186415  0.570000  0.350000  0.160000  0.280000  0.210000  0.440000
0.670000
186064  0.480000  0.330000  0.030000  0.350000  0.290000  0.720000
0.600000
38552   0.373498  0.345578  0.392939  0.339569  0.291659  0.477997
0.548875
149111  0.300000  0.290000  0.510000  0.250000  0.200000  0.510000
0.990000
...          ...       ...       ...       ...       ...       ...
...
31991   0.110000  0.510000  0.940000  0.090000  0.090000  0.700000
0.100000
25672   0.480000  0.480000  0.460000  0.710000  0.720000  0.710000
0.460000
81988   0.470000  0.480000  0.480000  0.480000  0.490000  0.550000
0.540000
97594   0.340000  0.690000  0.080000  0.060000  0.050000  0.750000
0.520000
53332   0.373498  0.345578  0.392939  0.339569  0.291659  0.477997
0.548875

             Q11      Q12      Q13       Q14       Q15       Q16
Q17  \
168265  0.650000  0.51000  1.00000  0.860000  0.040000  0.358732
1.000000
186415  0.730000  0.62000  0.88000  0.670000  0.480000  0.700000
0.690000
186064  0.510000  0.53000  0.41000  0.450000  0.060000  0.030000
0.480000
38552   0.586235  0.53572  0.47002  0.533462  0.395239  0.358732
0.537583
149111  0.670000  0.69000  0.46000  0.460000  0.430000  0.200000
0.860000
...          ...      ...      ...       ...       ...       ...
...
31991   0.340000  0.10000  0.10000  0.110000  0.110000  0.130000
0.550000
25672   0.480000  0.51000  0.51000  0.310000  0.960000  0.070000
0.510000
81988   0.550000  0.54000  0.53000  0.540000  0.530000  0.550000
0.550000
97594   0.570000  0.21000  0.15000  0.270000  0.110000  0.140000
0.750000
53332   0.586235  0.53572  0.47002  0.533462  0.395239  0.358732
```

0.537583

```
             Q18       Q19 AGE_GROUP  LIKE_ARTIST_1-10
LIKE_ARTIST_11-20  \
168265  0.040000  0.520000     36-50              0.0
0.0
186415  0.600000  0.300000     18-25              0.0
0.0
186064  0.340000  0.330000     26-35              0.0
0.0
38552   0.422153  0.412752     36-50              0.0
0.0
149111  0.220000  0.340000     36-50              0.0
0.0
...          ...       ...       ...              ...
...
31991   0.100000  0.110000     51-65              0.0
0.0
25672   0.080000  0.090000     51-65              0.0
0.0
81988   0.422153  0.412752     26-35              0.0
0.0
97594   0.180000  0.130000     51-65              0.0
0.0
53332   0.422153  0.412752     36-50              0.0
0.0

        LIKE_ARTIST_21-30  LIKE_ARTIST_31-40  LIKE_ARTIST_41-50  \
168265                0.0                0.0                0.0
186415                0.0                0.0                0.0
186064                0.0                0.0                0.0
38552                 0.0                0.0                0.0
149111                0.0                0.0                0.0
...                   ...                ...                ...
31991                 1.0                0.0                0.0
25672                 0.0                0.0                0.0
81988                 0.0                0.0                0.0
97594                 0.0                0.0                0.0
53332                 0.0                0.0                0.0

        LIKE_ARTIST_51-60  LIKE_ARTIST_61-70  LIKE_ARTIST_71-80  \
168265                0.0                0.0                0.0
186415                0.0                0.0                0.0
186064                0.0                0.0                0.0
38552                 0.0                0.0                0.0
149111                0.0                0.0                0.0
...                   ...                ...                ...
31991                 0.0                0.0                0.0
25672                 0.0                0.0                0.0
81988                 0.0                0.0                0.0
```

```
97594                 0.0                 0.0                 0.0
53332                 0.0                 0.0                 0.0

        LIKE_ARTIST_81-90  LIKE_ARTIST_91-100  LIKE_ARTIST_No
Answer  \
168265                0.0                  0.0                     1.0

186415                0.0                  0.0                     1.0

186064                0.0                  0.0                     1.0

38552                 0.0                  0.0                     1.0

149111                0.0                  0.0                     1.0

...                   ...                  ...                     ...

31991                 0.0                  0.0                     0.0

25672                 0.0                  0.0                     1.0

81988                 0.0                  0.0                     1.0

97594                 0.0                  0.0                     1.0

53332                 0.0                  0.0                     1.0


        GENDER_Female  GENDER_Male  GENDER_No Answer  \
168265            1.0          0.0                0.0
186415            0.0          1.0                0.0
186064            0.0          1.0                0.0
38552             0.0          0.0                1.0
149111            1.0          0.0                0.0
...               ...          ...                ...
31991             0.0          1.0                0.0
25672             1.0          0.0                0.0
81988             0.0          1.0                0.0
97594             0.0          1.0                0.0
53332             0.0          0.0                1.0

        WORKING_Employed 30+ hours a week  \
168265                                0.0
186415                                0.0
186064                                1.0
38552                                 0.0
149111                                0.0
...                                   ...
31991                                 0.0
```

```
25672                                    0.0
81988                                    0.0
97594                                    1.0
53332                                    0.0

        WORKING_Employed 8-29 hours per week  \
168265                                    0.0
186415                                    0.0
186064                                    0.0
38552                                     0.0
149111                                    1.0
...                                       ...
31991                                     0.0
25672                                     0.0
81988                                     0.0
97594                                     0.0
53332                                     0.0

        WORKING_Employed part-time less than 8 hours per week  \
168265                                                    0.0
186415                                                    0.0
186064                                                    0.0
38552                                                     0.0
149111                                                    0.0
...                                                       ...
31991                                                     0.0
25672                                                     0.0
81988                                                     0.0
97594                                                     0.0
53332                                                     0.0

        WORKING_Full-time housewife / househusband  WORKING_Full-time
student  \
168265                                          1.0
0.0
186415                                          0.0
1.0
186064                                          0.0
0.0
38552                                           0.0
0.0
149111                                          0.0
0.0
...                                             ...
...
31991                                           0.0
0.0
25672                                           0.0
0.0
81988                                           0.0
```

```
0.0
97594                                                0.0
0.0
53332                                                0.0
0.0

        WORKING_In unpaid employment (e.g. voluntary work)  WORKING_No
Answer  \
168265                                                    0.0
0.0
186415                                                    0.0
0.0
186064                                                    0.0
0.0
38552                                                     0.0
1.0
149111                                                    0.0
0.0
...                                                       ...
...
31991                                                     0.0
0.0
25672                                                     0.0
0.0
81988                                                     0.0
1.0
97594                                                     0.0
0.0
53332                                                     0.0
1.0

        WORKING_Other  WORKING_Part-time student  WORKING_Prefer not
to state  \
168265            0.0                       0.0
0.0
186415            0.0                       0.0
0.0
186064            0.0                       0.0
0.0
38552             0.0                       0.0
0.0
149111            0.0                       0.0
0.0
...               ...                       ...
...
31991             0.0                       0.0
0.0
25672             0.0                       0.0
0.0
81988             0.0                       0.0
```

0.0
97594   0.0   0.0
0.0
53332   0.0   0.0
0.0

        WORKING_Retired from full-time employment (30+ hours per week)
\
168265   0.0

186415   0.0

186064   0.0

38552   0.0

149111   0.0

...   ...

31991   0.0

25672   1.0

81988   0.0

97594   0.0

53332   0.0

        WORKING_Retired from self-employment WORKING_Self-employed \
168265   0.0   0.0
186415   0.0   0.0
186064   0.0   0.0
38552   0.0   0.0
149111   0.0   0.0
...   .
31991   0.0   1.0
25672   0.0   0.0
81988   0.0   0.0
97594   0.0   0.0
53332   0.0   0.0

        WORKING_Temporarily unemployed REGION_Centre REGION_Midlands
\
168265   0.0   0.0   1.0

186415   0.0   0.0   0.0

```
186064                              0.0            0.0              0.0

38552                               0.0            0.0              0.0

149111                              0.0            0.0              1.0

...                                 ...            ...              ...

31991                               0.0            0.0              0.0

25672                               0.0            0.0              0.0

81988                               0.0            0.0              0.0

97594                               0.0            0.0              0.0

53332                               0.0            0.0              0.0


        REGION_No Answer  REGION_North  REGION_Northern Ireland
REGION_South  \
168265               0.0           0.0                      0.0
0.0
186415               0.0           0.0                      0.0
1.0
186064               0.0           0.0                      0.0
1.0
38552                1.0           0.0                      0.0
0.0
149111               0.0           0.0                      0.0
0.0
...                  ...           ...                      ...
...
31991                0.0           1.0                      0.0
0.0
25672                0.0           1.0                      0.0
0.0
81988                0.0           1.0                      0.0
0.0
97594                0.0           1.0                      0.0
0.0
53332                1.0           0.0                      0.0
0.0

        LIST_OWN_0  LIST_OWN_0.5  LIST_OWN_1  LIST_OWN_11-14
LIST_OWN_15-19  \
168265         0.0           0.0         1.0             0.0
0.0
```

```
186415         0.0           0.0         0.0             0.0
0.0
186064         0.0           1.0         0.0             0.0
0.0
38552          0.0           0.0         0.0             0.0
0.0
149111         0.0           0.0         0.0             0.0
0.0
...            ...           ...         ...             ...
...
31991          1.0           0.0         0.0             0.0
0.0
25672          0.0           1.0         0.0             0.0
0.0
81988          0.0           0.0         0.0             0.0
0.0
97594          0.0           0.0         1.0             0.0
0.0
53332          0.0           0.0         0.0             0.0
0.0

        LIST_OWN_2  LIST_OWN_20 and plus  LIST_OWN_3-6  LIST_OWN_7-10
\
168265         0.0                   0.0           0.0            0.0

186415         1.0                   0.0           0.0            0.0

186064         0.0                   0.0           0.0            0.0

38552          0.0                   0.0           0.0            0.0

149111         1.0                   0.0           0.0            0.0

...            ...                   ...           ...            ...

31991          0.0                   0.0           0.0            0.0

25672          0.0                   0.0           0.0            0.0

81988          0.0                   0.0           1.0            0.0

97594          0.0                   0.0           0.0            0.0

53332          0.0                   0.0           0.0            0.0


        LIST_OWN_No Answer  LIST_BACK_0  LIST_BACK_0.5  LIST_BACK_1  \
168265                 0.0          0.0            0.0          1.0
186415                 0.0          0.0            0.0          0.0
```

```
186064                  0.0          1.0            0.0          0.0
38552                   1.0          0.0            0.0          0.0
149111                  0.0          0.0            0.0          0.0
...                     ...          ...            ...          ...
31991                   0.0          1.0            0.0          0.0
25672                   0.0          0.0            0.0          0.0
81988                   0.0          0.0            0.0          0.0
97594                   0.0          0.0            1.0          0.0
53332                   1.0          0.0            0.0          0.0

        LIST_BACK_11-14  LIST_BACK_15-19  LIST_BACK_2  LIST_BACK_20
and plus  \
168265              0.0              0.0          0.0
0.0
186415              0.0              0.0          0.0
0.0
186064              0.0              0.0          0.0
0.0
38552               0.0              0.0          0.0
0.0
149111              0.0              0.0          0.0
0.0
...                 ...              ...          ...
...
31991               0.0              0.0          0.0
0.0
25672               0.0              0.0          0.0
0.0
81988               0.0              0.0          0.0
0.0
97594               0.0              0.0          0.0
0.0
53332               0.0              0.0          0.0
0.0

        LIST_BACK_3-6  LIST_BACK_7-10  LIST_BACK_No Answer
AGE_GROUP_13-17  \
168265            0.0             0.0                  0.0
0.0
186415            1.0             0.0                  0.0
0.0
186064            0.0             0.0                  0.0
0.0
38552             0.0             0.0                  1.0
0.0
149111            1.0             0.0                  0.0
0.0
...               ...             ...                  ...
...
31991             0.0             0.0                  0.0
```

```
0.0
25672             0.0             1.0                   0.0
0.0
81988             1.0             0.0                   0.0
0.0
97594             0.0             0.0                   0.0
0.0
53332             0.0             0.0                   1.0
0.0

        AGE_GROUP_18-25  AGE_GROUP_26-35  AGE_GROUP_36-50
AGE_GROUP_51-65  \
168265              0.0              0.0              1.0
0.0
186415              1.0              0.0              0.0
0.0
186064              0.0              1.0              0.0
0.0
38552               0.0              0.0              1.0
0.0
149111              0.0              0.0              1.0
0.0
...                 ...              ...              ...
...
31991               0.0              0.0              0.0
1.0
25672               0.0              0.0              0.0
1.0
81988               0.0              1.0              0.0
0.0
97594               0.0              0.0              0.0
1.0
53332               0.0              0.0              1.0
0.0

        AGE_GROUP_older than 65
168265                      0.0
186415                      0.0
186064                      0.0
38552                       0.0
149111                      0.0
...                         ...
31991                       0.0
25672                       0.0
81988                       0.0
97594                       0.0
53332                       0.0

[150952 rows x 94 columns]
```

```
# Saving to Disk

print('training_inputs:', training_inputs.shape)
print('training_targets:', training_targets.shape)
print('validation_inputs:', validation_inputs.shape)
print('validation_targets:', validation_targets.shape)
print('test_inputs:', test_inputs.shape)
```

```
training_inputs: (150952, 94)
training_targets: (150952,)
validation_inputs: (37738, 94)
validation_targets: (37738,)
test_inputs: (125794, 94)
```

```
!pip install pyarrow --quiet

training_inputs.to_parquet('training_inputs.parquet')
validation_inputs.to_parquet('validation_inputs.parquet')
test_inputs.to_parquet('test_inputs.parquet')

pd.DataFrame(training_targets).to_parquet('training_targets.parquet')
pd.DataFrame(validation_targets).to_parquet('validation_targets.parque
t')
```

Getting Data Back

```
training_inputs = pd.read_parquet('training_inputs.parquet')
validation_inputs = pd.read_parquet('validation_inputs.parquet')
test_inputs = pd.read_parquet('test_inputs.parquet')

training_targets = pd.read_parquet('training_targets.parquet')
[target_col]
validation_targets = pd.read_parquet('validation_targets.parquet')
[target_col]

print('training_inputs:', training_inputs.shape)
print('training_targets:', training_targets.shape)
print('validation_inputs:', validation_inputs.shape)
print('validation_targets:', validation_targets.shape)
print('test_inputs:', test_inputs.shape)
```

```
training_inputs: (150952, 94)
training_targets: (150952,)
validation_inputs: (37738, 94)
validation_targets: (37738,)
test_inputs: (125794, 94)
```

Starting Modeling

```
X_training = training_inputs[numeric_cols + encoded_cols]
```

```
X_validation = validation_inputs[numeric_cols + encoded_cols]

X_test = test_inputs[numeric_cols + encoded_cols]
```

Training

```
from xgboost import XGBRegressor

model = XGBRegressor(n_jobs=0)

model.fit(X_training, training_targets)
```

```
[11:34:29] WARNING: /workspace/src/objective/regression_obj.cu:152:
reg:linear is now deprecated in favor of reg:squarederror.
```

```
XGBRegressor(n_jobs=0)
```

```
prediction = model.predict(X_training)

from sklearn.metrics import mean_squared_error

def rmse(a, b):
  return mean_squared_error(a, b, squared=False)

rmse(prediction, training_targets)
```

```
15.97886769351449
```

```
impt_df = pd.DataFrame({'feature': X_training.columns,
'importance': model.feature_importances_}).sort_values('importance',
ascending=False)

impt_df.head(10)
```

	feature	importance
6	words_score	0.467572
5	OWN_ARTIST_MUSIC	0.083540
4	HEARD_OF	0.042305
3	Time	0.029398
18	Q11	0.026738
14	Q7	0.025278
36	LIKE_ARTIST_91-100	0.022223
19	Q12	0.020269
23	Q16	0.018554
15	Q8	0.016989

Hyperparametre Tuning

```
def test_params(**params):
  model = XGBRegressor(n_jobs=-1, **params)
  model.fit(X_training, training_targets)
  training_rmse = rmse(model.predict(X_training), training_targets)
  validation_rmse = rmse(model.predict(X_validation),
```

```
validation_targets)
  print('Training RMSE: {}, Validation RMSE: {}'.format(training_rmse,
validation_rmse))

test_params(n_estimators=100)

[11:35:23] WARNING: /workspace/src/objective/regression_obj.cu:152:
reg:linear is now deprecated in favor of reg:squarederror.
Training RMSE: 15.97886769351449, Validation RMSE: 15.909823636844786

test_params(n_estimators=200)

[11:35:52] WARNING: /workspace/src/objective/regression_obj.cu:152:
reg:linear is now deprecated in favor of reg:squarederror.
Training RMSE: 15.77299591949312, Validation RMSE: 15.741147592019022

test_params(n_estimators=400)

[11:36:48] WARNING: /workspace/src/objective/regression_obj.cu:152:
reg:linear is now deprecated in favor of reg:squarederror.
Training RMSE: 15.526375941069244, Validation RMSE: 15.569182444072737

test_params(n_estimators=800)

[11:38:41] WARNING: /workspace/src/objective/regression_obj.cu:152:
reg:linear is now deprecated in favor of reg:squarederror.
Training RMSE: 15.193733025548545, Validation RMSE: 15.370873997573677
```

Tree depth & Learning rate

```
test_params(n_estimators=175, max_depth=8, learning_rate=0.3)

[11:42:28] WARNING: /workspace/src/objective/regression_obj.cu:152:
reg:linear is now deprecated in favor of reg:squarederror.
Training RMSE: 10.777195811333101, Validation RMSE: 14.35447977385776

test_params(n_estimators=175, max_depth=8, learning_rate=0.2)

[11:44:56] WARNING: /workspace/src/objective/regression_obj.cu:152:
reg:linear is now deprecated in favor of reg:squarederror.
Training RMSE: 11.705383800905652, Validation RMSE: 14.38046099437259

test_params(booster='gblinear', n_estimators=400)

[11:47:14] WARNING: /workspace/src/objective/regression_obj.cu:152:
reg:linear is now deprecated in favor of reg:squarederror.
Training RMSE: 21.690106222953066, Validation RMSE: 21.679318734157885

test_params(n_estimators=500, max_depth=9, learning_rate=0.15)

[11:48:11] WARNING: /workspace/src/objective/regression_obj.cu:152:
reg:linear is now deprecated in favor of reg:squarederror.
Training RMSE: 8.170350841099964, Validation RMSE: 14.031425644970993
```

```
test_params(n_estimators=1000, max_depth=10, learning_rate=0.10,
subsample=0.9, colsample_bytree=0.7)

[11:56:22] WARNING: /workspace/src/objective/regression_obj.cu:152:
reg:linear is now deprecated in favor of reg:squarederror.
Training RMSE: 5.912378909481084, Validation RMSE: 14.06289233604193

from sklearn.model_selection import KFold

def train_and_evaluate(X_train_k, Y_train_k, X_val_k, Y_val_k,
**params):
  model = XGBRegressor(n_jobs=-1, **params)
  model.fit(X_train_k, Y_train_k)
  train_rmse = rmse(model.predict(X_train_k), Y_train_k)
  val_rmse = rmse(model.predict(X_val_k), Y_val_k)
  return model, train_rmse, val_rmse

kfold = KFold(n_splits=5)

models = []

for train_idxs, val_idxs in kfold.split(X_training):
  X_train_k, Y_train_k = X_training.iloc[train_idxs],
training_targets.iloc[train_idxs]
  X_val_k, Y_val_k = X_training.iloc[val_idxs],
training_targets.iloc[val_idxs]
  model, train_rmse, val_rmse = train_and_evaluate(X_train_k,
Y_train_k, X_val_k, Y_val_k, n_estimators=500, max_depth=9,
learning_rate=0.10, subsample=0.9, colsample_bytree=0.7)
  models.append(model)
  print('Train RMSE: {}, Validation RMSE: {}'.format(train_rmse,
val_rmse))

[12:11:43] WARNING: /workspace/src/objective/regression_obj.cu:152:
reg:linear is now deprecated in favor of reg:squarederror.
Train RMSE: 8.87027198507148, Validation RMSE: 14.309372055537372
[12:17:14] WARNING: /workspace/src/objective/regression_obj.cu:152:
reg:linear is now deprecated in favor of reg:squarederror.
Train RMSE: 8.930045003488926, Validation RMSE: 14.326546756947248
[12:22:31] WARNING: /workspace/src/objective/regression_obj.cu:152:
reg:linear is now deprecated in favor of reg:squarederror.
Train RMSE: 8.883624648870502, Validation RMSE: 14.325172790349075
[12:27:49] WARNING: /workspace/src/objective/regression_obj.cu:152:
reg:linear is now deprecated in favor of reg:squarederror.
Train RMSE: 8.964643709238226, Validation RMSE: 14.379924947602829
[12:33:01] WARNING: /workspace/src/objective/regression_obj.cu:152:
reg:linear is now deprecated in favor of reg:squarederror.
Train RMSE: 8.926999279240505, Validation RMSE: 14.242177838373042

def predict_avg(models, inputs):
    return np.mean([model.predict(inputs) for model in models],
axis=0)
```

```
preds_kfold = predict_avg(models, X_validation)
rmse(preds_kfold, validation_targets)
```

```
13.852739686869784
```

```
test_preds = predict_avg(models, X_test)
```

Final Answer

```
test_preds.shape
```

```
(125794,)
```

Model 2

RandomForestRegressor

```
from sklearn.ensemble import RandomForestRegressor

model_randomForestRegressor = RandomForestRegressor()

model_randomForestRegressor.fit(X_training, training_targets)
```

```
RandomForestRegressor()
```

```
def test_params(**params):
    model = RandomForestRegressor(random_state=42, n_jobs=-1,
**params).fit(X_training, training_targets)
    return model.score(X_training, training_targets),
model.score(X_validation, validation_targets)
```

Hyperparameter Tuning

```
test_params(max_depth=100, max_leaf_nodes=2**4)
```

```
(0.44986499029102844, 0.4560022614096648)
```

```
test_params(max_depth=400, max_leaf_nodes=2**10)
```

```
(0.5880712862974407, 0.5376298342989975)
```

```
test_params(max_depth=600, max_leaf_nodes=2**15)
```

```
(0.9241145573397733, 0.5908789148855196)
```

```
test_params(max_depth=1000, max_leaf_nodes=2**25)
```

```
(0.9419978303117578, 0.5872249941906067)
```

Now we can see that the optimum value occurs at `max_depth=600` & `max_leaf_nodes=2**15`

Now we will use these values to predict the performance of RFR.

Performance of RFR

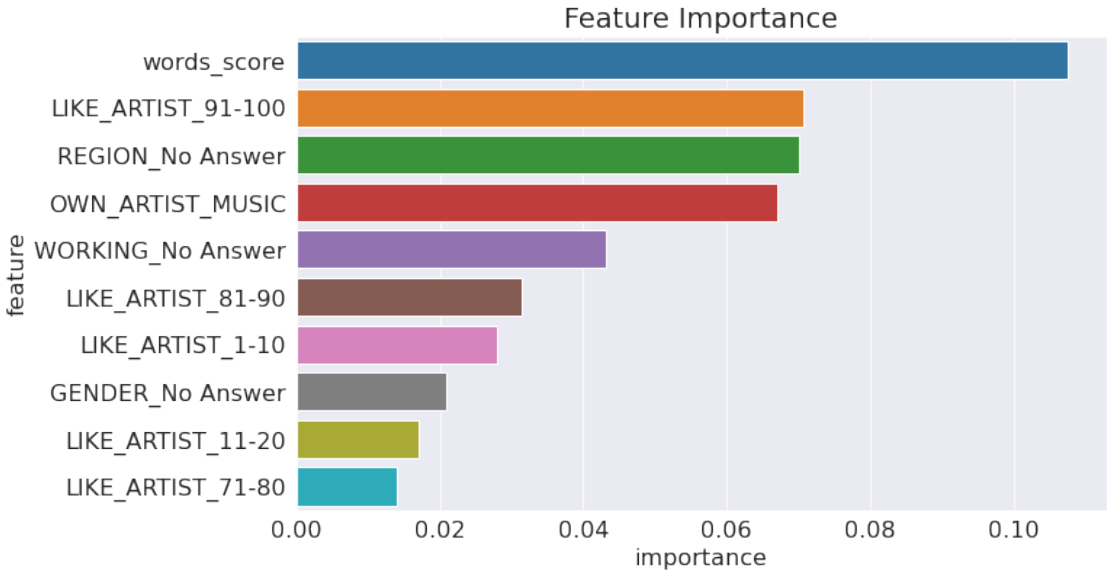
```
preds_randomForestRegressor =
model_randomForestRegressor.predict(X_validation)
rmse(preds_randomForestRegressor, validation_targets)
```

```
14.504533421368059
```

Performance of RFR is less than XGBR So, we will continue with XGBR Model

Importance of columns

```
importance_df = pd.DataFrame({'feature': X_training.columns,
'importance': model.feature_importances_}).sort_values('importance',
ascending=False)

plt.figure(figsize=(10,6))
plt.title('Feature Importance')
sns.barplot(data=importance_df.head(10), x='importance', y='feature');
```



Saving the model

```
import joblib

song_recommendation_ml = {
    'model': models,
    'imputer': imputer,
    'scaler': scaler,
    'encoder': encoder,
    'input_cols': input_cols,
    'target_cols': target_col,
    'numeric_cols': numeric_cols,
    'categorical_cols': categorical_cols,
```

```
    'encoded_cols': encoded_cols
}

joblib.dump(song_recommendation_ml, 'song_recommendation_ml')

['song_recommendation_ml']

['song_recommendation_ml']

['song_recommendation_ml']
```

Conclusion

I downloded this dataset from kaggle. Then after I imported the requried python libraries. Now, I started cleaning the dataset like deleting the rows in which data is missing or substituting the average value of the column in the missing data. Then to get the insights from the columns of the dataset, I started to make the visualizations of the dataset. After I got th einsights from the dataset and the relationship betweeen the columns of the dataset I started to make the model. I used XGBoosta and RFR models. After checking the performances of both the models, I had come to a conclusion that XGBoost model had high performance than RFR. So, at last I used the XGBoost model to predict the values of dataset.

References and Future Work

References: The websites that I found useful during this project work are Scikit-learn, Stackoverflow, W3schools, GFG, and many more.

- GFG
- scikit-learn
- GFG
- Stack overflow
- Medium

Future work

Now, I will continue on this project, by adding the songs data and selecting the recommended song for the listner from the dataset. In the dataset of the song we have to differentiate the songs by the lyrics in the song, by the lyrics of the song we can is it a sad, happy or romatic song. By the words in the lyrics of the song we can recommed the type of song that the listerner wants.