data-cleanup

December 31, 2024

[30]:
```
import pandas as pd
import logging
# Configure logging
logging.basicConfig(level=logging.INFO)
```

[31]:
```
# Read the CSV file from the data/raw folder
input_file_path = 'data/raw/Spotify_Youtube.csv'
df = pd.read_csv(input_file_path)
```

[32]:
```
df.head()
```

[32]:

	Id	Artist	Url_spotify	\
0	0	Gorillaz	https://open.spotify.com/artist/3AA28KZvwAUcZu…	
1	1	Gorillaz	https://open.spotify.com/artist/3AA28KZvwAUcZu…	
2	2	Gorillaz	https://open.spotify.com/artist/3AA28KZvwAUcZu…	
3	3	Gorillaz	https://open.spotify.com/artist/3AA28KZvwAUcZu…	
4	4	Gorillaz	https://open.spotify.com/artist/3AA28KZvwAUcZu…	

	Track	\
0	Feel Good Inc.	
1	Rhinestone Eyes	
2	New Gold (feat. Tame Impala and Bootie Brown)	
3	On Melancholy Hill	
4	Clint Eastwood	

	Album	Album_type	\
0	Demon Days	album	
1	Plastic Beach	album	
2	New Gold (feat. Tame Impala and Bootie Brown)	single	
3	Plastic Beach	album	
4	Gorillaz	album	

	Uri	Danceability	Energy	Key	…	\
0	spotify:track:0d28khcov6AiegSCpG5TuT	0.818	0.705	6.0	…	
1	spotify:track:1foMv2HQwfQ2vntFf9HFeG	0.676	0.703	8.0	…	
2	spotify:track:64dLd6rVqDLtkXFYrEUHIU	0.695	0.923	1.0	…	
3	spotify:track:0q6LuUqGLUiCPP1cbdwFs3	0.689	0.739	2.0	…	
4	spotify:track:7yMiX7n9SBvadzox8T5jzT	0.663	0.694	10.0	…	

```
                                    Url_youtube  \
0  https://www.youtube.com/watch?v=HyHNuVaZJ-k
1  https://www.youtube.com/watch?v=yYDmaexVHic
2  https://www.youtube.com/watch?v=qJa-VFwPpYA
3  https://www.youtube.com/watch?v=04mfKJWDSzI
4  https://www.youtube.com/watch?v=1V_xRb0x9aw

                                                Title   Channel        Views  \
0         Gorillaz - Feel Good Inc. (Official Video)  Gorillaz  693555221.0
1  Gorillaz - Rhinestone Eyes [Storyboard Film] (…  Gorillaz   72011645.0
2  Gorillaz - New Gold ft. Tame Impala & Bootie B…  Gorillaz    8435055.0
3     Gorillaz - On Melancholy Hill (Official Video)  Gorillaz  211754952.0
4         Gorillaz - Clint Eastwood (Official Video)  Gorillaz  618480958.0

       Likes  Comments                                          Description  \
0  6220896.0  169907.0  Official HD Video for Gorillaz' fantastic trac…
1  1079128.0   31003.0  The official video for Gorillaz - Rhinestone E…
2   282142.0    7399.0  Gorillaz - New Gold ft. Tame Impala & Bootie B…
3  1788577.0   55229.0  Follow Gorillaz online:\nhttp://gorillaz.com \…
4  6197318.0  155930.0  The official music video for Gorillaz - Clint …

   Licensed official_video       Stream
0      True           True  1.040235e+09
1      True           True  3.100837e+08
2      True           True  6.306347e+07
3      True           True  4.346636e+08
4      True           True  6.172597e+08

[5 rows x 28 columns]
```

[33]:
```
# Remove columens has no use to the analysis and keeps big data
df = df.drop(columns=['Url_spotify','Url_youtube','Description'])

# Save the modified CSV file to the data/processed folder
output_file_path = 'data/processed/Spotify_Youtube.csv'
df.to_csv(output_file_path, index=False)

logging.info(f"Modified CSV saved to {output_file_path}")

df.head()
```

```
INFO:root:Modified CSV saved to data/processed/Spotify_Youtube.csv
```

[33]:
```
   Id    Artist                                              Track  \
0   0  Gorillaz                                    Feel Good Inc.
1   1  Gorillaz                                   Rhinestone Eyes
2   2  Gorillaz  New Gold (feat. Tame Impala and Bootie Brown)
```

```
3   3  Gorillaz                                On Melancholy Hill
4   4  Gorillaz                                    Clint Eastwood

                                              Album Album_type  \
0                                        Demon Days      album
1                                     Plastic Beach      album
2  New Gold (feat. Tame Impala and Bootie Brown)     single
3                                     Plastic Beach      album
4                                          Gorillaz      album

                                    Uri  Danceability  Energy   Key  Loudness  \
0  spotify:track:0d28khcov6AiegSCpG5TuT         0.818   0.705   6.0    -6.679
1  spotify:track:1foMv2HQwfQ2vntFf9HFeG         0.676   0.703   8.0    -5.815
2  spotify:track:64dLd6rVqDLtkXFYrEUHIU         0.695   0.923   1.0    -3.930
3  spotify:track:0q6LuUqGLUiCPP1cbdwFs3         0.689   0.739   2.0    -5.810
4  spotify:track:7yMiX7n9SBvadzox8T5jzT         0.663   0.694  10.0    -8.627

   …    Tempo  Duration_ms  \
0  …  138.559     222640.0
1  …   92.761     200173.0
2  …  108.014     215150.0
3  …  120.423     233867.0
4  …  167.953     340920.0

                                                Title   Channel        Views  \
0        Gorillaz - Feel Good Inc. (Official Video)  Gorillaz  693555221.0
1  Gorillaz - Rhinestone Eyes [Storyboard Film] (…  Gorillaz   72011645.0
2  Gorillaz - New Gold ft. Tame Impala & Bootie B…  Gorillaz    8435055.0
3    Gorillaz - On Melancholy Hill (Official Video)  Gorillaz  211754952.0
4        Gorillaz - Clint Eastwood (Official Video)  Gorillaz  618480958.0

       Likes  Comments Licensed official_video        Stream
0  6220896.0  169907.0     True           True  1.040235e+09
1  1079128.0   31003.0     True           True  3.100837e+08
2   282142.0    7399.0     True           True  6.306347e+07
3  1788577.0   55229.0     True           True  4.346636e+08
4  6197318.0  155930.0     True           True  6.172597e+08

[5 rows x 25 columns]
```

[35]:
```python
# Read the CSV file from the data/raw folder with specified encoding
most_streamed_file_path = 'data/raw/Most Streamed Spotify Songs 2024.csv'
most_streamed_df = pd.read_csv(most_streamed_file_path, encoding='ISO-8859-1')

# Display the first few rows of the dataframe
most_streamed_df.head()
```

[35]:
```
                        Track                    Album Name          Artist  \
0         MILLION DOLLAR BABY  Million Dollar Baby - Single   Tommy Richman
1                 Not Like Us                 Not Like Us  Kendrick Lamar
2  i like the way you kiss me  I like the way you kiss me         Artemas
3                     Flowers            Flowers - Single     Miley Cyrus
4                     Houdini                    Houdini          Eminem

  Release Date          ISRC All Time Rank  Track Score Spotify Streams  \
0    4/26/2024  QM24S2402528             1        725.4     390,470,936
1     5/4/2024  USUG12400910             2        545.9     323,703,884
2    3/19/2024  QZJ842400387             3        538.4     601,309,283
3    1/12/2023  USSM12209777             4        444.9   2,031,280,633
4    5/31/2024  USUG12403398             5        423.3     107,034,922

  Spotify Playlist Count Spotify Playlist Reach  …  SiriusXM Spins  \
0                 30,716            196,631,588  …             684
1                 28,113            174,597,137  …               3
2                 54,331            211,607,669  …             536
3                269,802            136,569,078  …           2,182
4                  7,223            151,469,874  …               1

  Deezer Playlist Count Deezer Playlist Reach Amazon Playlist Count  \
0                  62.0            17,598,718                 114.0
1                  67.0            10,422,430                 111.0
2                 136.0            36,321,847                 172.0
3                 264.0            24,684,248                 210.0
4                  82.0            17,660,624                 105.0

  Pandora Streams Pandora Track Stations Soundcloud Streams  Shazam Counts  \
0      18,004,655                 22,931          4,818,457      2,669,262
1       7,780,028                 28,444          6,623,075      1,118,279
2       5,022,621                  5,639          7,208,651      5,285,340
3     190,260,277                203,384                NaN     11,822,942
4       4,493,884                  7,006            207,179        457,017

  TIDAL Popularity Explicit Track
0              NaN              0
1              NaN              1
2              NaN              0
3              NaN              0
4              NaN              1

[5 rows x 29 columns]
```

[36]:
```
spotify_2000_file_path = 'data/raw/Spotify-2000.csv'
spotify_2000_df = pd.read_csv(spotify_2000_file_path)
spotify_2000_df.head()
```

[36]:

	Index	Title	Artist	Top Genre	\
0	1	Sunrise	Norah Jones	adult standards	
1	2	Black Night	Deep Purple	album rock	
2	3	Clint Eastwood	Gorillaz	alternative hip hop	
3	4	The Pretender	Foo Fighters	alternative metal	
4	5	Waitin' On A Sunny Day	Bruce Springsteen	classic rock	

	Year	Beats Per Minute (BPM)	Energy	Danceability	Loudness (dB)	\
0	2004	157	30	53	-14	
1	2000	135	79	50	-11	
2	2001	168	69	66	-9	
3	2007	173	96	43	-4	
4	2002	106	82	58	-5	

	Liveness	Valence	Length (Duration)	Acousticness	Speechiness	Popularity
0	11	68	201	94	3	71
1	17	81	207	17	7	39
2	7	52	341	2	17	69
3	3	37	269	0	4	76
4	10	87	256	1	3	59

[ ]: