databricks Ajay_BDS_Assignment_Notebook 2023-12-24 20:37:56

(https://databricks.com)

```
file_path = "/FileStore/train.csv"
```

```
df = spark.read.option("header", "true").csv(file_path)
```

```
df.show()
```

```
+----------+--------------------+--------------------+--------------------+--------------------+----------+-----------+------
--+------------+------+---+--------+----+-----------+------------+---------------+--------+-------+-------+--------------+--
---------+
|Unnamed: 0|            track_id|             artists|          album_name|          track_name|popularity|duration_ms|explic
it|danceability|energy|key|loudness|mode|speechiness|acousticness|instrumentalness|liveness|valence|  tempo|time_signature|tr
ack_genre|
+----------+--------------------+--------------------+--------------------+--------------------+----------+-----------+------
--+------------+------+---+--------+----+-----------+------------+---------------+--------+-------+-------+--------------+--
---------+
|         0|5SuOikwiRyPMVoIQD...|         Gen Hoshino|              Comedy|              Comedy|        73|     230666|   Fal
se|       0.676| 0.461|  1|  -6.746|   0|      0.143|      0.0322|        1.01e-06|   0.358|  0.715| 87.917|             4|
acoustic|
|         1|4qPNDBW1i3p13qLCt...|        Ben Woodward|    Ghost (Acoustic)|    Ghost - Acoustic|        55|     149610|   Fal
se|        0.42| 0.166|  1| -17.235|   1|     0.0763|       0.924|        5.56e-06|   0.101|  0.267| 77.489|             4|
acoustic|
|         2|1iJBSr7s7jYXzM8EG...|Ingrid Michaelson...|      To Begin Again|      To Begin Again|        57|     210826|   Fal
se|       0.438| 0.359|  0|  -9.734|   1|     0.0557|        0.21|             0.0|   0.117|   0.12| 76.332|             4|
acoustic|
|         3|6lfxq3CG4xtTiEg7o...|        Kina Grannis|Crazy Rich Asians...|Can't Help Fallin...|        71|     201933|   Fal
se|       0.266|0.0596|  0| -18.515|   1|     0.0363|       0.905|        7.07e-05|   0.132|  0.143| 181.74|             3|
acoustic|
```

```
#Find out the most popular artists:
df.createOrReplaceTempView("spotify_data")

popular_artists = spark.sql("""
SELECT artists, COUNT(track_id) as total_tracks, AVG(popularity) as avg_popularity
FROM spotify_data
GROUP BY artists
ORDER BY avg_popularity DESC
LIMIT 10
""")
popular_artists.show()
```

```
+--------------------+------------+--------------+
|             artists|total_tracks|avg_popularity|
+--------------------+------------+--------------+
|Sam Smith;Kim Petras|           2|         100.0|
|    Bizarrap;Quevedo|           1|          99.0|
|       Manuel Turizo|           4|          98.0|
|Bad Bunny;Chencho...|           4|          97.0|
|Bad Bunny;Bomba E...|           4|          94.5|
|                Joji|           1|          94.0|
|             Beyoncé|           1|          93.0|
|   Rema;Selena Gomez|           1|          92.0|
|        Harry Styles|           3|          92.0|
|     Drake;21 Savage|           1|          91.0|
+--------------------+------------+--------------+
```

```
#Analyze songs attributes associated with  popularity >80:
song_attributes = spark.sql("""
SELECT AVG(danceability) as avg_danceability, AVG(energy) as avg_energy, AVG(loudness) as avg_loudness, AVG(valence) as
avg_valence, AVG(tempo) as avg_tempo
FROM spotify_data
WHERE popularity > 80
""")
song_attributes.show()
```

```
+------------------+------------------+------------------+------------------+------------------+
|  avg_danceability|        avg_energy|      avg_loudness|       avg_valence|         avg_tempo|
+------------------+------------------+------------------+------------------+------------------+
|0.6581111111111113|0.6794260167714885|-6.127298742138366|0.5079800838574422|118.42675471698118|
+------------------+------------------+------------------+------------------+------------------+
```

```
  # average danceability, energy, and loudness for the top 5 popular songs
  top_5_popular_songs = spark.sql("""
  SELECT track_name, artists, album_name, popularity, danceability, energy, loudness
  FROM spotify_data
  WHERE popularity > 80
  ORDER BY popularity DESC
  LIMIT 5
  """)
  top_5_popular_songs.show()

  # average danceability, energy, and loudness for the top 5 popular songs
  least_5_popular_songs = spark.sql("""
  SELECT track_name, artists, album_name, popularity, danceability, energy, loudness
  FROM spotify_data
  WHERE popularity < 15
  ORDER BY popularity ASC
  LIMIT 5
  """)
  least_5_popular_songs.show()


  # Calculate the average danceability, energy, and loudness for the top 5 popular songs
  avg_top_danceability = top_5_popular_songs.agg({"danceability": "avg"}).collect()[0][0]
  avg_top_energy = top_5_popular_songs.agg({"energy": "avg"}).collect()[0][0]
  avg_top_loudness = top_5_popular_songs.agg({"loudness": "avg"}).collect()[0][0]

  # Calculate the average danceability, energy, and loudness for the least 5 popular songs
  avg_least_danceability = least_5_popular_songs.agg({"danceability": "avg"}).collect()[0][0]
  avg_least_energy = least_5_popular_songs.agg({"energy": "avg"}).collect()[0][0]
  avg_least_loudness = least_5_popular_songs.agg({"loudness": "avg"}).collect()[0][0]

  # Compare
  print("Average Danceability - Top: {}, Least: {}".format(avg_top_danceability, avg_least_danceability))
  print("Average Energy - Top: {}, Least: {}".format(avg_top_energy, avg_least_energy))
  print("Average Loudness - Top: {}, Least: {}".format(avg_top_loudness, avg_least_loudness))
```

```
+--------------------+--------------------+--------------------+----------+------------+------+--------+
|          track_name|             artists|          album_name|popularity|danceability|energy|loudness|
+--------------------+--------------------+--------------------+----------+------------+------+--------+
|Quevedo: Bzrp Mus...|    Bizarrap;Quevedo|Quevedo: Bzrp Mus...|        99|       0.621| 0.782|  -5.548|
|     I'm Good (Blue)|David Guetta;Bebe...|     I'm Good (Blue)|        98|       0.561| 0.965|  -3.673|
|          La Bachata|       Manuel Turizo|          La Bachata|        98|       0.835| 0.679|  -5.329|
|     I'm Good (Blue)|David Guetta;Bebe...|     I'm Good (Blue)|        98|       0.561| 0.965|  -3.673|
|          La Bachata|       Manuel Turizo|          La Bachata|        98|       0.835| 0.679|  -5.329|
+--------------------+--------------------+--------------------+----------+------------+------+--------+

+--------------------+--------------------+--------------------+----------+------------+------+--------+
|          track_name|             artists|          album_name|popularity|danceability|energy|loudness|
+--------------------+--------------------+--------------------+----------+------------+------+--------+
|Psyjack - DJ Snea...|Sidney Charles;DJ...|House Music Risin...|         0|       0.805| 0.914|  -5.181|
|         Dead Inside|            Chimaira|Groove Metal Root...|         0|       0.384| 0.982|  -3.132|
|   Winter Wonderland|          Jason Mraz|      Christmas Time|         0|        0.62| 0.309|  -9.209|
|If You Think This...|Destroy Rebuild U...|Post-Millennial N...|         0|       0.597|  0.97|  -6.303|
|          De Colores|Ismael Rivera Y S...|Era Salsa para Ba...|         0|       0.492| 0.765|  -4.661|
+--------------------+--------------------+--------------------+----------+------------+------+--------+

Average Danceability - Top: 0.6826, Least: 0.6498000000000002
```

```
+------------+
|avg_duration|
+------------+
|         2.8|
+------------+

+------------+
```

```
|avg_duration|
+------------+
|        10.5|
+------------+

Average Duration of Top 5 Popular Songs: 2.8
Average Duration of Least 5 Popular Songs: 10.5
```