Exploratory Data Analysis using the Top 50 Spotify Songs

Taylor Saunders DAT301 Fall 2023

Introduction

This report displays an Exploratory Data Analysis of the top 50 spotify songs in 2019

Data set sourced from Kaggle: https://www.kaggle.com/datasets/leonardopena/top50spotify2019/data (https://www.kaggle.com/datasets/leonardopena/top50spotify2019/data)

What is Spotify?

Spotify is a music streaming platform that was established in 2006. Since then, spotify has grown to a size of about 489 million users. Due to the platforms popularity, it is a good tool to view what songs/artists are popular world wide and to analyze music trends.

Beginning with the Data

First, I am loading in the libraries I will be using for analysis and visualization.

In [90]:
```
import numpy as np
import pandas as pd
#To visualize the data
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore') #supress warning messages packages may produ
ce
```

Next, to begin our data anaylsis I loaded in the spreadsheet containing the information about the top 50 songs on spotify. The dataset will be referred to as "mydf" for the duration of this report.

In [5]:
```
mydf = pd.read_csv("top50spotifyutf.csv")
```

Data Cleaning

The first step to a successful exploratory data analysis is data cleaning. This ensures that the numbers we produce in our report are accurate and relevant to the reality of the data presented.

The first step I will perform is to remove any NA values that may be present.

In [7]:
```
mydf.dropna()
```

Out[7]:

	Unnamed: 0	Track.Name	Artist.Name	Genre	Beats.Per.Minute	Energy	Danceability	L
0	1	Señorita	Shawn Mendes	canadian pop	117	55	76	
1	2	China	Anuel AA	reggaeton flow	105	81	79	
2	3	boyfriend (with Social House)	Ariana Grande	dance pop	190	80	40	
3	4	Beautiful People (feat. Khalid)	Ed Sheeran	pop	93	65	64	
4	5	Goodbyes (Feat. Young Thug)	Post Malone	dfw rap	150	65	58	
5	6	I Don't Care (with Justin Bieber)	Ed Sheeran	pop	102	68	80	
6	7	Ransom	Lil Tecca	trap music	180	64	75	
7	8	How Do You Sleep?	Sam Smith	pop	111	68	48	
8	9	Old Town Road - Remix	Lil Nas X	country rap	136	62	88	
9	10	bad guy	Billie Eilish	electropop	135	43	70	
10	11	Callaita	Bad Bunny	reggaeton	176	62	61	
11	12	Loco Contigo (feat. J. Balvin & Tyga)	DJ Snake	dance pop	96	71	82	
12	13	Someone You Loved	Lewis Capaldi	pop	110	41	50	
13	14	Otro Trago - Remix	Sech	panamanian pop	176	79	73	
14	15	Money In The Grave (Drake ft. Rick Ross)	Drake	canadian hip hop	101	50	83	
15	16	No Guidance (feat. Drake)	Chris Brown	dance pop	93	45	70	
16	17	LA CANCIÓN	J Balvin	latin	176	65	75	
17	18	Sunflower - Spider-Man: Into the Spider-Verse	Post Malone	dfw rap	90	48	76	
18	19	Lalala	Y2K	canadian hip hop	130	39	84	
19	20	Truth Hurts	Lizzo	escape room	158	62	72	
20	21	Piece Of Your Heart	MEDUZA	pop house	124	74	68	

	Unnamed: 0	Track.Name	Artist.Name	Genre	Beats.Per.Minute	Energy	Danceability	L
21	22	Panini	Lil Nas X	country rap	154	59	70	
22	23	No Me Conoce - Remix	Jhay Cortez	reggaeton flow	92	79	81	
23	24	Soltera - Remix	Lunay	latin	92	78	80	
24	25	bad guy (with Justin Bieber)	Billie Eilish	electropop	135	45	67	
25	26	If I Can't Have You	Shawn Mendes	canadian pop	124	82	69	
26	27	Dance Monkey	Tones and I	australian pop	98	59	82	
27	28	It's You	Ali Gatie	canadian hip hop	96	46	73	
28	29	Con Calma	Daddy Yankee	latin	94	86	74	
29	30	QUE PRETENDES	J Balvin	latin	93	79	64	
30	31	Takeaway	The Chainsmokers	edm	85	51	29	
31	32	7 rings	Ariana Grande	dance pop	140	32	78	
32	33	0.958333333	Maluma	reggaeton	96	71	78	
33	34	The London (feat. J. Cole & Travis Scott)	Young Thug	atl hip hop	98	59	80	
34	35	Never Really Over	Katy Perry	dance pop	100	88	77	
35	36	Summer Days (feat. Macklemore & Patrick Stump ...	Martin Garrix	big room	114	72	66	
36	37	Otro Trago	Sech	panamanian pop	176	70	75	
37	38	Antisocial (with Travis Scott)	Ed Sheeran	pop	152	82	72	
38	39	Sucker	Jonas Brothers	boy band	138	73	84	
39	40	fuck, i'm lonely (with Anne-Marie) - from “13 ...	Lauv	dance pop	95	56	81	
40	41	Higher Love	Kygo	edm	104	68	69	
41	42	You Need To Calm Down	Taylor Swift	dance pop	85	68	77	

	Unnamed: 0	Track.Name	Artist.Name	Genre	Beats.Per.Minute	Energy	Danceability	L
42	43	Shallow	Lady Gaga	dance pop	96	39	57	
43	44	Talk	Khalid	pop	136	40	90	
44	45	Con Altura	ROSALÍA	r&b en espanol	98	69	88	
45	46	One Thing Right	Marshmello	brostep	88	62	66	
46	47	Te Robaré	Nicky Jam	latin	176	75	67	
47	48	Happier	Marshmello	brostep	100	79	69	
48	49	Call You Mine	The Chainsmokers	edm	104	70	59	
49	50	Cross Me (feat. Chance the Rapper & PnB Rock)	Ed Sheeran	pop	95	79	75	

In [9]:
```
mydf.shape
```

Out[9]:
```
(50, 14)
```

As we can see, we still have 50 rows of data. Therefore, there were no NA values in our dataset.

To further our analysis, I will begin to get us comfortable with the data by performing various steps to determine classes we are dealing with and an idea of what kind of genres are in our dataset.

In [12]:
```
mydf.head()
```

Out[12]:

	Unnamed: 0	Track.Name	Artist.Name	Genre	Beats.Per.Minute	Energy	Danceability	Loudne
0	1	Señorita	Shawn Mendes	canadian pop	117	55	76	
1	2	China	Anuel AA	reggaeton flow	105	81	79	
2	3	boyfriend (with Social House)	Ariana Grande	dance pop	190	80	40	
3	4	Beautiful People (feat. Khalid)	Ed Sheeran	pop	93	65	64	
4	5	Goodbyes (Feat. Young Thug)	Post Malone	dfw rap	150	65	58	

In [13]:
```
mydf.tail()
```

Out[13]:

	Unnamed: 0	Track.Name	Artist.Name	Genre	Beats.Per.Minute	Energy	Danceability	Loudne
45	46	One Thing Right	Marshmello	brostep	88	62	66	
46	47	Te Robaré	Nicky Jam	latin	176	75	67	
47	48	Happier	Marshmello	brostep	100	79	69	
48	49	Call You Mine	The Chainsmokers	edm	104	70	59	
49	50	Cross Me (feat. Chance the Rapper & PnB Rock)	Ed Sheeran	pop	95	79	75	

As we can see, the columns we are dealing with are:

Genre

Beats per Minute

Energy

Danceability

Loudness (dB)

Liveness

Valence

Length

Acousticness

Speechiness

Popularity

These are all unique values that will help us determine correlation between the top 50 songs.

In [15]:
```
mydf.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 50 entries, 0 to 49
Data columns (total 14 columns):
 #   Column            Non-Null Count  Dtype 
---  ------            --------------  ----- 
 0   Unnamed: 0        50 non-null     int64 
 1   Track.Name        50 non-null     object
 2   Artist.Name       50 non-null     object
 3   Genre             50 non-null     object
 4   Beats.Per.Minute  50 non-null     int64 
 5   Energy            50 non-null     int64 
 6   Danceability      50 non-null     int64 
 7   Loudness..dB..    50 non-null     int64 
 8   Liveness          50 non-null     int64 
 9   Valence.          50 non-null     int64 
 10  Length.           50 non-null     int64 
 11  Acousticness..    50 non-null     int64 
 12  Speechiness.      50 non-null     int64 
 13  Popularity        50 non-null     int64 
dtypes: int64(11), object(3)
memory usage: 5.6+ KB
```

The two types of variables we are dealing with are objects and integers.

Data Analysis

In [80]:
```
mydf.describe()
```

Out[80]:

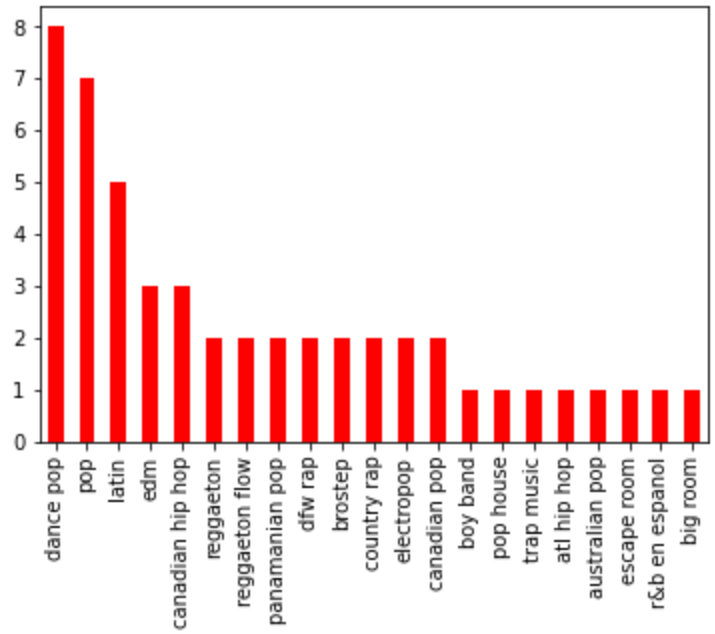
	Unnamed: 0	Beats.Per.Minute	Energy	Danceability	Loudness..dB..	Liveness	Valence.
count	50.00000	50.000000	50.000000	50.00000	50.000000	50.000000	50.000000
mean	25.50000	120.060000	64.060000	71.38000	-5.660000	14.660000	54.600000
std	14.57738	30.898392	14.231913	11.92988	2.056448	11.118306	22.336024
min	1.00000	85.000000	32.000000	29.00000	-11.000000	5.000000	10.000000
25%	13.25000	96.000000	55.250000	67.00000	-6.750000	8.000000	38.250000
50%	25.50000	104.500000	66.500000	73.50000	-6.000000	11.000000	55.500000
75%	37.75000	137.500000	74.750000	79.75000	-4.000000	15.750000	69.500000
max	50.00000	190.000000	88.000000	90.00000	-2.000000	58.000000	95.000000

In [61]:
```
mydf['Genre'].value_counts()
```

Out[61]:
```
dance pop           8
pop                 7
latin               5
edm                 3
canadian hip hop    3
reggaeton           2
reggaeton flow      2
panamanian pop      2
dfw rap             2
brostep             2
country rap         2
electropop          2
canadian pop        2
boy band            1
pop house           1
trap music          1
atl hip hop         1
australian pop      1
escape room         1
r&b en espanol      1
big room            1
Name: Genre, dtype: int64
```

In [63]:
```
plt.figure()
mydf['Genre'].value_counts().plot(kind = 'bar', color = 'r')
```

Out[63]:
```
<AxesSubplot:>
```



The most common genre is dance pop, closely followed by pop.

In [21]:
```
mydf['Artist.Name'].value_counts()
```

Out[21]:
```
Ed Sheeran          4
J Balvin            2
The Chainsmokers    2
Ariana Grande       2
Marshmello          2
Lil Nas X           2
Billie Eilish       2
Shawn Mendes        2
Sech                2
Post Malone         2
Maluma              1
Katy Perry          1
Bad Bunny           1
Drake               1
DJ Snake            1
Lizzo               1
Martin Garrix       1
Young Thug          1
Taylor Swift        1
Sam Smith           1
Lewis Capaldi       1
Lauv                1
Y2K                 1
Ali Gatie           1
Daddy Yankee        1
Jhay Cortez         1
Jonas Brothers      1
Tones and I         1
Lunay               1
Lady Gaga           1
Chris Brown         1
ROSALÍA             1
MEDUZA              1
Kygo                1
Khalid              1
Nicky Jam           1
Lil Tecca           1
Anuel AA            1
Name: Artist.Name, dtype: int64
```

Ed Sheeran had the most popular songs, amounting to 4 out of the 50 top songs.

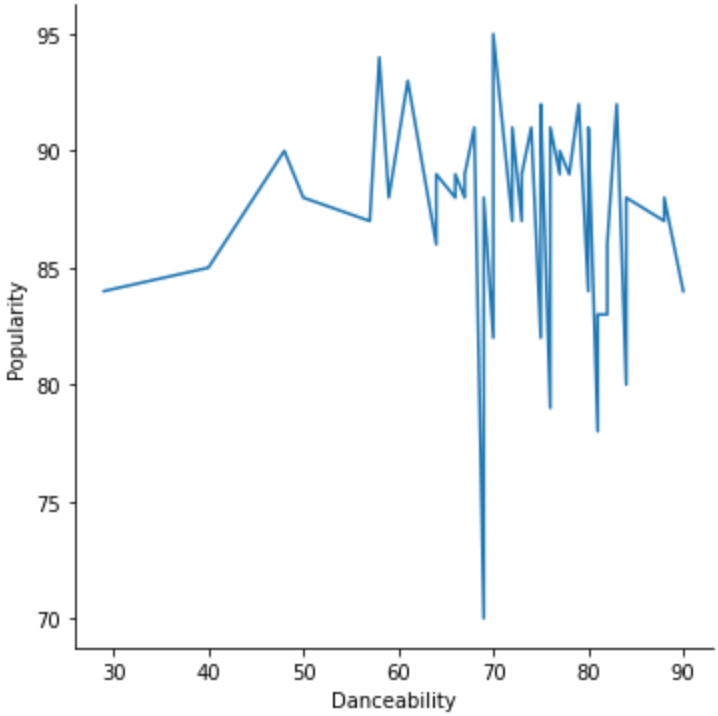
Data Visualization

Now, lets begin to visualize our data to derive conclusions to the following research questions:

1. Is there a connection betweeen danceability and popularity?
2. What music genre was the most popular in 2019 on Spotify?
3. What is the correlation between danceability and pop genres

In [71]:
```
sns.relplot(
    data=mydf, kind="line",
    x="Danceability", y="Popularity",
    estimator=None,
)
```

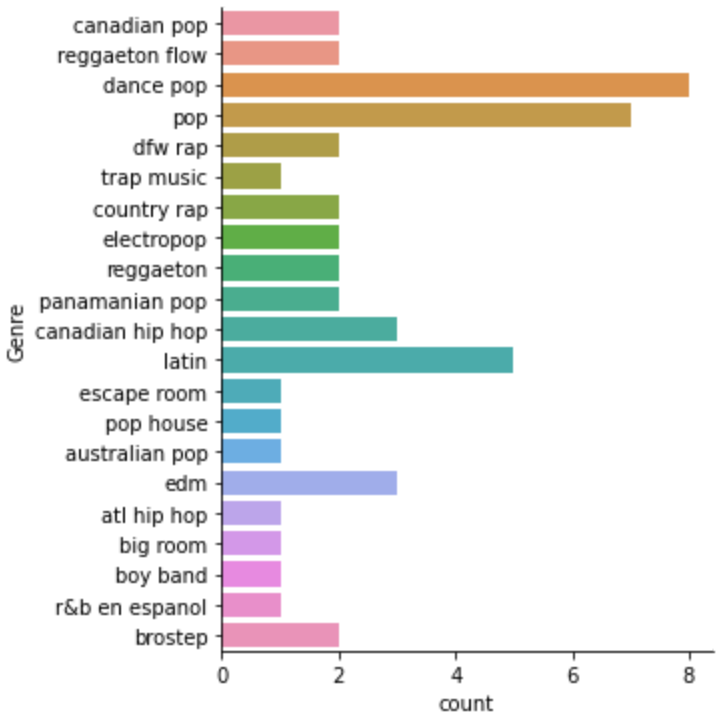
Out[71]:
```
<seaborn.axisgrid.FacetGrid at 0x7f376826ab20>
```



There does not appear to be a clear correlation between danceability and popularity based on this graph.

In [35]:
```
sns.catplot(data= mydf, y='Genre', kind= 'count')
```

Out[35]:
```
<seaborn.axisgrid.FacetGrid at 0x7f37820b7850>
```



Based on this plot, we can see the most popular genres are dance pop and pop. Because of this analysis, it feeds into the next question:

What is the correlation between pop genres and danceability?

In [73]:
```
for i in mydf['Genre']:
    if 'pop' in i:
        mydf['Genre'] = mydf['Genre'].replace(i, 'pop')
```

In [69]:
```
mydf['Genre'].unique()
```

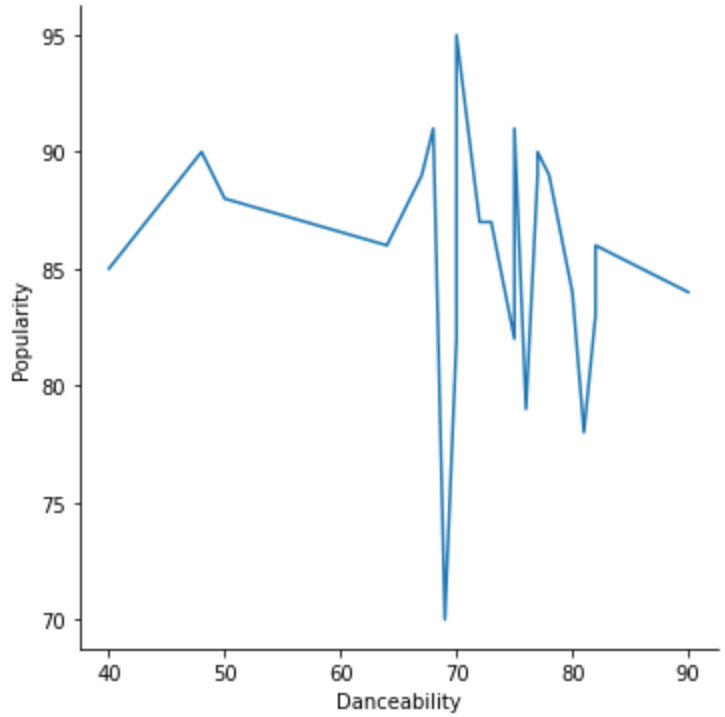
Out[69]:
```
array(['pop', 'reggaeton flow', 'dfw rap', 'trap music', 'country rap',
       'reggaeton', 'canadian hip hop', 'latin', 'escape room', 'edm',
       'atl hip hop', 'big room', 'boy band', 'r&b en espanol', 'brostep'],
      dtype=object)
```

After grouping all genres including "pop", we can begin simple visualization.

In [76]:
```
mypopdf = mydf[mydf['Genre'] == 'pop']
```

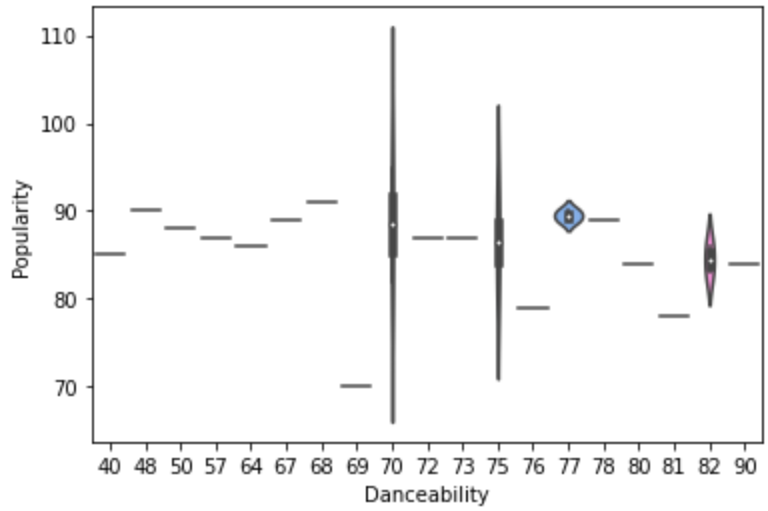
In [78]:
```
sns.relplot(
    data=mypopdf, kind="line",
    x="Danceability", y="Popularity",
    estimator=None,
)
```

Out[78]: <seaborn.axisgrid.FacetGrid at 0x7f3768177580>



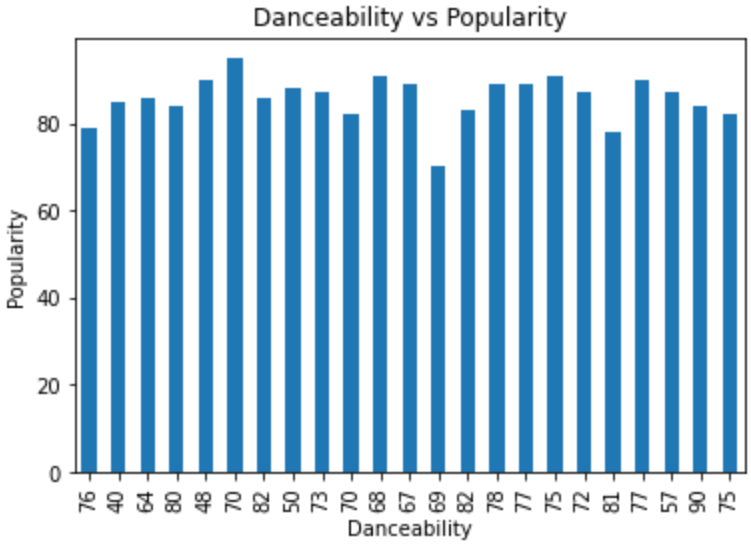
This looks very similar to our plot above. Lets try a different kind of plot to see if we can get a clearer result.

In [83]:
```
sns.violinplot(x='Danceability', y='Popularity', data=mypopdf)
plt.show()
```



This isnt very clear either - lets try one more!

In [85]:
```
mypopdf.plot(kind='bar', x='Danceability', y='Popularity', legend=False)
plt.title('Danceability vs Popularity')
plt.xlabel('Danceability')
plt.ylabel('Popularity')
plt.show()
```



After three tests, I have come to the conclusion that danceability has no correlation to popularity for either pop-specific songs or all genres in this dataframe.

Conclusion

Based on my research, I concluded that danceability does not affect popularity (in the top 50 songs) and pop genres are the most popular genre. Additionally, there is a wide variety of genres and artists in the top 50; the artist with the most songs in top 50 (Ed Sheernan) only amounts to 4 of the songs. This suggests users of Spotify are very diverse, and more research could be done using music statistics from spotify as music is an interesting demographic that may correlate to different backgrounds.

Possible Future Work
Further analysis can be done to determine what popular artists are doing to succeed on Spotify's platform, and it could be replecated by smaller bands to determine if replicating beats per minute and dB levels can create higher desire for music.