IE 6200 Probability and Statistics

Project Report on Analysis of Musical Attributes of Spotify Dataset

Content

- Objective---1
- Part 1---1
  - 1.1Data set info and attribute range-------------------------2
  - 1.2 The trend over the years table and graphs-------------3
- Part 2- Correlation between attributes-------------------------4
  - 2.1. Analysis of the data set-----------------------------------4
- Part 3- Analysis of most popular songs -----------------------5
  - 3.1 Defining the range for most popular songs------------6
  - 3.2 Hypothesis testing for popular song attributes with the entire dataset---5
    - 3.2.1. Hypothesis testing 1-------------------------------5
    - 3.2.2. Hypothesis testing 2-------------------------------5
  - 3.3. Statistical Analysis of Most Popular Song------------6
    - 3.3.1. calculate the statistical values of attributes and graphical analysis--6
    - 3.3.2. Determining the attribute range for popular songs---7
- Conclusion---9
- Reference---10

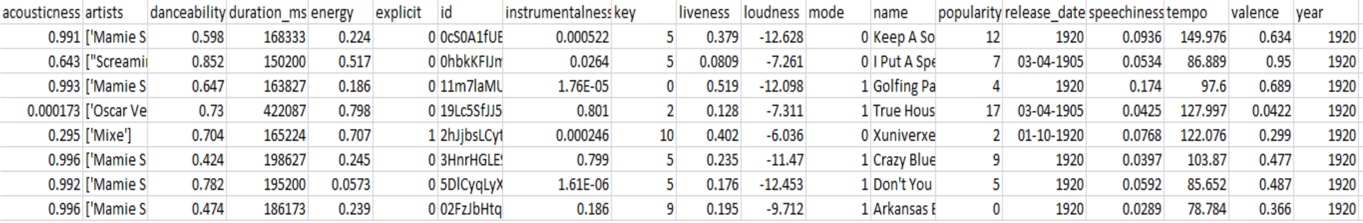
Objective- We will use a sizable dataset from Spotify for our study to spot trends and determine whether there is a set pattern or standardized attributes for producing hit songs. The used dataset has a little more than 170k rows and several different attributes as columns. To get at our results, we will perform an Exploratory Data Analysis (EDA) on a specific Spotify music dataset.

Part 1:

1.1. Data set info and attribute range-

Data set link -https://jovian.ai/zhangxm963/spotify-dataframe/v/39/files?filename=spotify-dataset-19212020-160k-tracks/data.csv
The initial dataset has the following key attributes:
Range Index: 174389 entries,
Data columns (total 19 columns):

- acousticness (Ranges from 0 to 1): A confidence measure from 0.0 to 1.0 of whether the track is acoustic.
- danceability (Ranges from 0 to 1): Danceability describes how suitable a track is for dancing.
- energy (Ranges from 0 to 1): Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity.
- duration_ms (Integer typically ranging from 200k to 300k)
- instrumentalness (Ranges from 0 to 1): Predicts whether a track contains no vocals. Values above 0.5 are intended to represent instrumental tracks.
- valence (Ranges from 0 to 1):
- popularity (Ranges from 0 to 100)
- tempo (Float typically ranging from 50 to 150)
- liveness (Ranges from 0 to 1): Detects the presence of an audience in the recording. A value above 0.8 provides a strong likelihood that the track is live.
- loudness (Float typically ranging from -60 to 0):
- speechiness (Ranges from 0 to 1)
- year (Ranges from 1921 to 2020) Dummy:
- mode (0 = Minor, 1 = Major)
- explicit (0 = No explicit content, 1 = Explicit content) Categorical:
- key (All keys on octave encoded as values ranging from 0 to 11, starting on C as 0, C# as 1, and so on…)
- artists (List of artists mentioned)
- release_date (Date of release mostly in yyyy-mm-dd format, however precision of date may vary)
- name (Name of the song)



acousticness	artists	danceability	duration_ms	energy	explicit	id	instrumentalness	key	liveness	loudness	mode	name	popularity	release_date	speechiness	tempo	valence	year
0.991	['Mamie S	0.598	168333	0.224	0	0cS0A1fUE	0.000522	5	0.379	-12.628	0	Keep A So	12	1920	0.0936	149.976	0.634	1920
0.643	["Screami	0.852	150200	0.517	0	0hbkKFIJm	0.0264	5	0.0809	-7.261	0	I Put A Spe	7	03-04-1905	0.0534	86.889	0.95	1920
0.993	['Mamie S	0.647	163827	0.186	0	11m7laMU	1.76E-05	0	0.519	-12.098	1	Golfing Pa	4	1920	0.174	97.6	0.689	1920
0.000173	['Oscar Ve	0.73	422087	0.798	0	19Lc5SfJJ5	0.801	2	0.128	-7.311	1	True Hous	17	03-04-1905	0.0425	127.997	0.0422	1920
0.295	['Mixe']	0.704	165224	0.707	1	2hJjbsLCyt	0.000246	10	0.402	-6.036	0	Xuniverxe	2	01-10-1920	0.0768	122.076	0.299	1920
0.996	['Mamie S	0.424	198627	0.245	0	3HnrHGLE!	0.799	5	0.235	-11.47	1	Crazy Blue	9	1920	0.0397	103.87	0.477	1920
0.992	['Mamie S	0.782	195200	0.0573	0	5DlCyqLyX	1.61E-06	5	0.176	-12.453	1	Don't You	5	1920	0.0592	85.652	0.487	1920
0.996	['Mamie S	0.474	186173	0.239	0	02FzJbHtq	0.186	9	0.195	-9.712	1	Arkansas E	0	1920	0.0289	78.784	0.366	1920

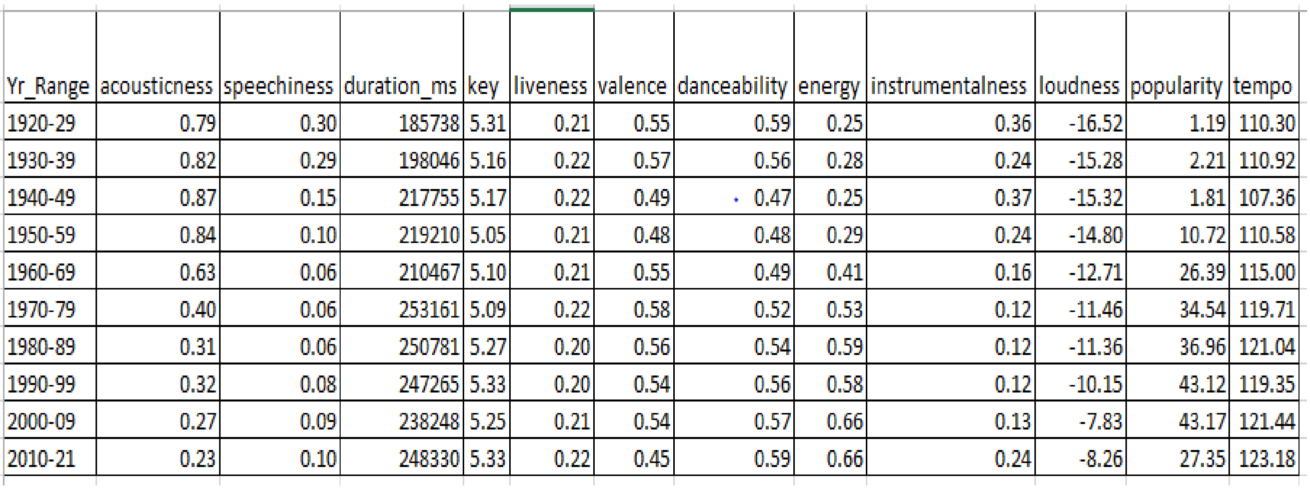
Data cleaning- However, since the – **user id, name, artist, and release date** column won't be that useful for our analysis, we can drop it.

1.2. The trend over the years table and graphs

Distribution of song attributes over the years-

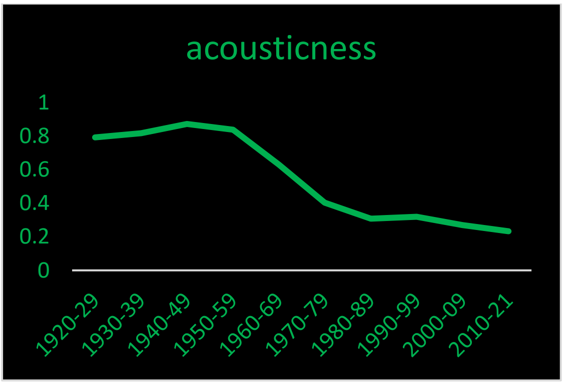
We divide all years from 1920 to 2021 into 10 sets with 10 years each and calculated mean values for each set for plotting the line graph. Below is the table of values and line graph for the same.

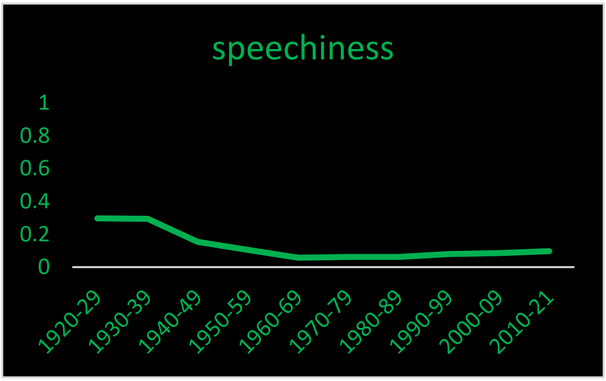
Table for mean values of attributes for the year range



Yr_Range	acousticness	speechiness	duration_ms	key	liveness	valence	danceability	energy	instrumentalness	loudness	popularity	tempo
1920-29	0.79	0.30	185738	5.31	0.21	0.55	0.59	0.25	0.36	-16.52	1.19	110.30
1930-39	0.82	0.29	198046	5.16	0.22	0.57	0.56	0.28	0.24	-15.28	2.21	110.92
1940-49	0.87	0.15	217755	5.17	0.22	0.49	0.47	0.25	0.37	-15.32	1.81	107.36
1950-59	0.84	0.10	219210	5.05	0.21	0.48	0.48	0.29	0.24	-14.80	10.72	110.58
1960-69	0.63	0.06	210467	5.10	0.21	0.55	0.49	0.41	0.16	-12.71	26.39	115.00
1970-79	0.40	0.06	253161	5.09	0.22	0.58	0.52	0.53	0.12	-11.46	34.54	119.71
1980-89	0.31	0.06	250781	5.27	0.20	0.56	0.54	0.59	0.12	-11.36	36.96	121.04
1990-99	0.32	0.08	247265	5.33	0.20	0.54	0.56	0.58	0.12	-10.15	43.12	119.35
2000-09	0.27	0.09	238248	5.25	0.21	0.54	0.57	0.66	0.13	-7.83	43.17	121.44
2010-21	0.23	0.10	248330	5.33	0.22	0.45	0.59	0.66	0.24	-8.26	27.35	123.18

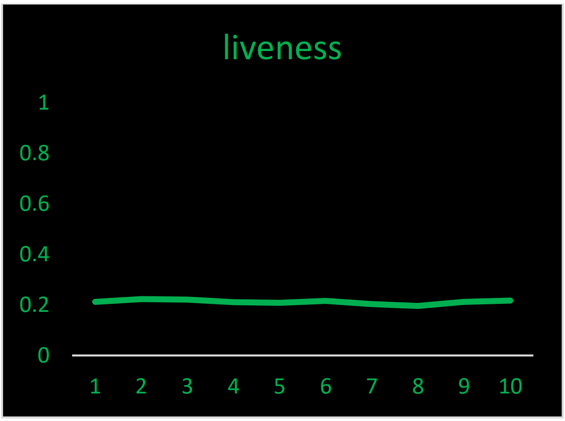
Line plots for the attributes over the year ranges.

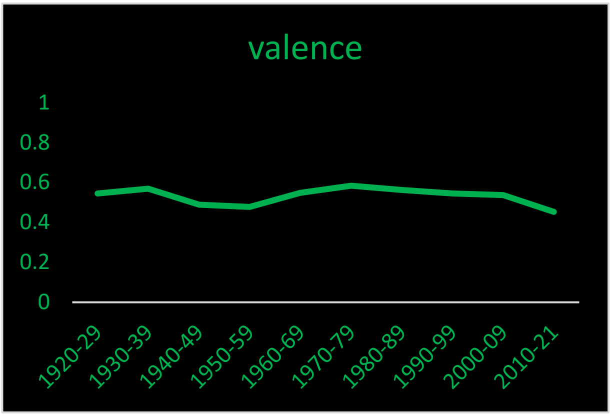




Acousticness - It has been observed that Acousticness gradually rose through the years 1920-1940 and from 1950 drastic decrease in the trend till 2021

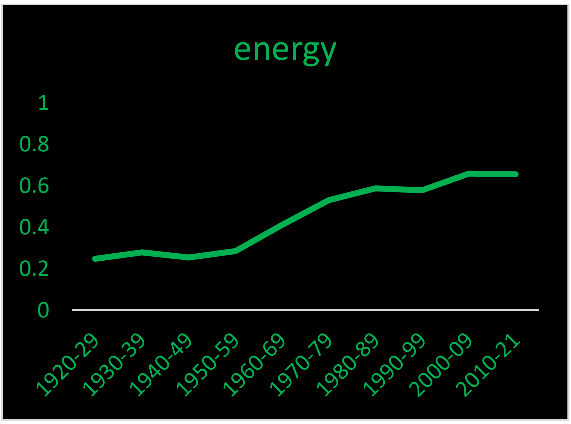
Speechiness - It can be observed from the graph that speechiness has been decreasing since the year 1939 and it became nearly constant after 1959.

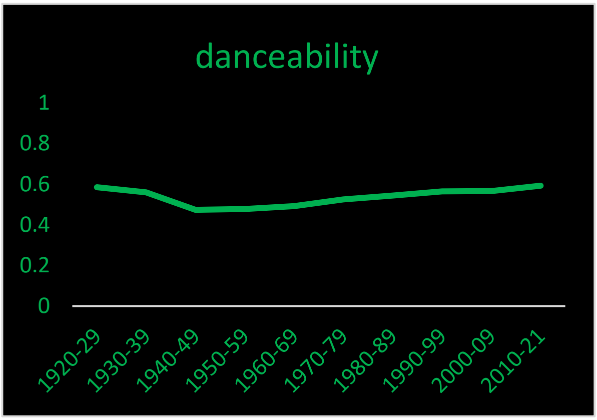




Liveness - It has been constant over the year with no significant change over the years.

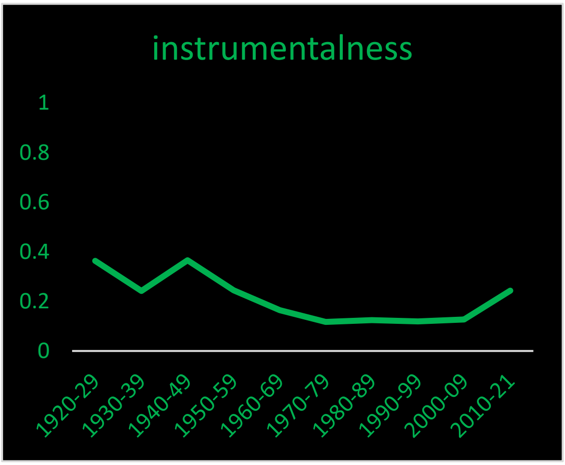
Valence- It has been constant more over the years where a slight dip can be observed in 1939

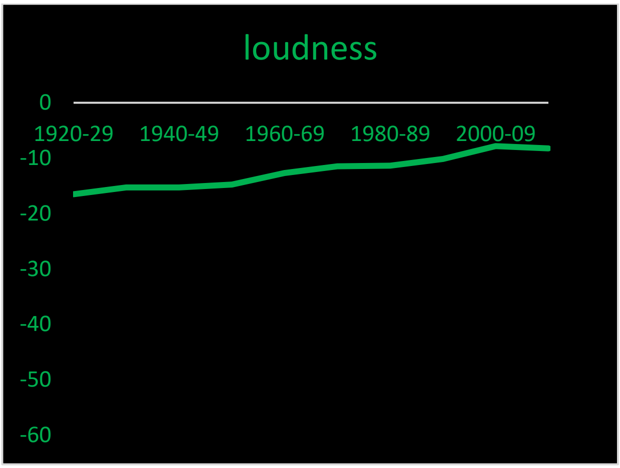




Energy - The attribute energy has been constant till 19549 and then there has been a drastic increase till 2021

Danceability - It has been constant over the years and a slight dip can be observed in 1940-1949





Instrumentalness -From the graph, it can be observed that it has declined in the first 2 decades then a sudden rise in 1940-49 and then it has gradually decreased.

Loudness - It can be observed that loudness has been steadily increasing over the years.

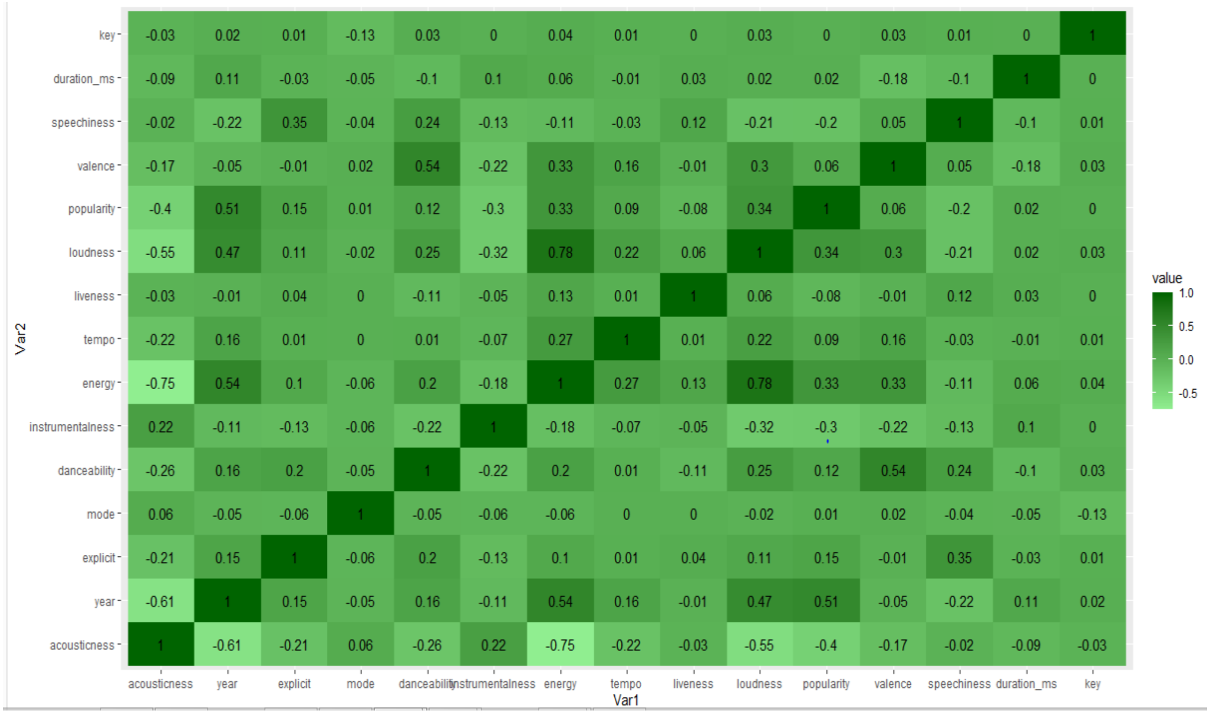
Part 2- Correlation between attributes

2.1. Analysis of the data set

We plotted a correlation heat map to understand how each attribute correlates with every other attribute.

The attributes taken into consideration are -

track attributes = [“acousticness”, “year”, “explicit”, “mode”, “danceability”, “instrumentalness”, “energy”, “tempo”, “liveness”, “popularity”, “valence”, “speechiness”, “duration_ms”, “key”]



Correlation Range Table-

0.0 - 0.2	0.2 - 0.4	0.4 - 0.6	0.6 - 0.8	0.8 - 1.0
Very Weak	Weak	Moderate	Strong	Very Strong

Inference from heat map-

Attribute 1	**Attribute 2**	**Correlation**	**Correlation Range**	**Correlation Strength**
Loudness	Energy	0.78	0.6 - 0.8	Positive High Correlation
Valence	Danceability	0.54	0.4 - 0.6	Moderate Correlation
Loudness	Instrumentalness	-0.32	0.2 - 0.4	Negative Low Correlation
Loudness	Acousticness	-0.55	0.4 - 0.6	Negative Moderate Correlation
Energy	Acousticness	-0.75	0.6 - 0.8	Negative High Correlation

From the heat map, it is evident that some of the attributes have a positive correlation between them and some have a negative correlation. Here, Positive Correlation indicates that as one attribute is increasing another is decreasing and negative correlation indicates that while one attribute is increasing other is decreasing.

Part 3: Analysis of most popular songs

3.1. Defining the range for most popular songs.

To analyze popular songs, we have to first define the range of popularity attributes for which we can consider the song popular. To do that we took the top 1% of songs (1743 songs) and considered their popularity range. We found that those songs have a popularity index above 75.

P (popularity >=75) = 0.0104

3.2. Hypothesis testing for popular song attributes with the entire dataset

After defining the popularity index range for popular songs, we divided the dataset for popular songs. So, we have 2 datasets now,

i) All songs listed on Spotify
ii) Most Popular songs (popularity index >=75)

Further, we use hypothesis testing to determine the relationship between the mean of two datasets using the sample mean and variance of samples from both datasets, assuming the population means and variance are unknown.

3.2.1. Hypothesis testing 1:

Is the population mean of acousticness for all listed songs more than popular songs?

Using R, a random sample of 200 from each dataset is taken. Their mean and variance are calculated. As the sample size is above 30, we can use the Z distribution. Stating hypothesis as below:

H0: $\mu1 < \mu2$: $\mu1 - \mu2 < 0$

H1: $\mu1 > \mu2$: $\mu1 - \mu2 > 0$

It’s a right-tailed test.

Calculating test statistic: Zcal = 6.24 and for 99% confidence Z(critical) = 2.326.

As Zcal > Z(critical), we reject the null hypothesis. So, with 99% confidence we can conclude that mean of acousticness for all listed songs is more than popular songs i.e. $\mu1 > \mu2$.

3.2.2. Hypothesis testing 2:

Is the population mean of danceability for all popular songs lesser than popular songs?

Here an also random sample of 200 from each dataset is taken. Their mean and variance are calculated. As the sample size is above 30, we used the Z distribution. Stating hypothesis as below:

H0: $\mu1 > \mu2$: $\mu1 - \mu2 > 0$

H1: $\mu1 < \mu2$: $\mu1 - \mu2 < 0$

It’s a left-tailed test.

Calculating test statistic: Zcal = - 7.38 and for 99% confidence Z (critical) = -2.326.

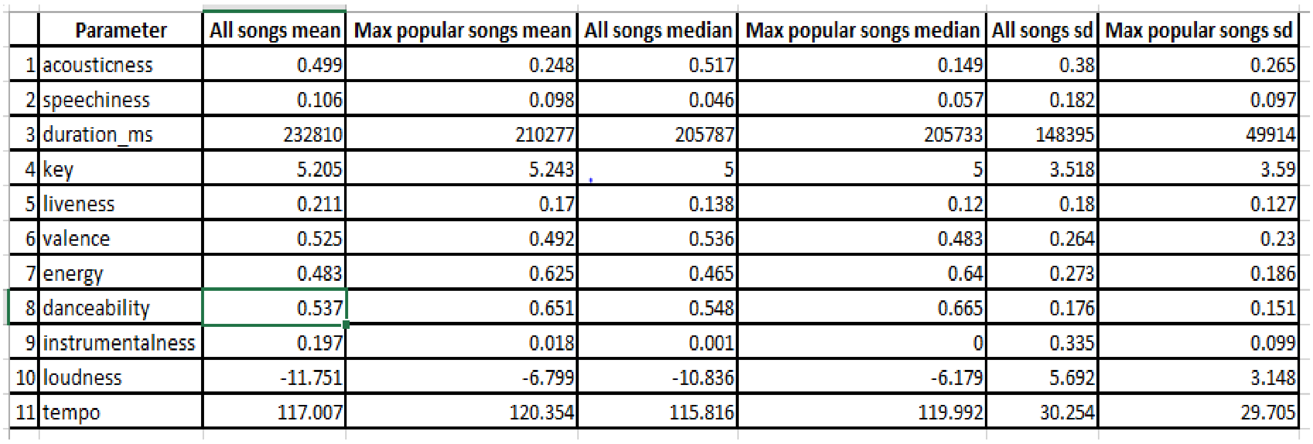
As Zcal < Z (critical), we reject the null hypothesis. So, with 99% confidence, we can conclude that mean of danceability for all listed songs is lesser than most popular songs i.e. µ1 < µ2.

3.3. Statistical Analysis of Most Popular Songs

3.3.1. calculate the statistical values of attributes and graphical analysis

In this way, we can do hypothesis testing for all attributes. However, using R we can calculate the population means directly to check the relationship between the attribute means of all listed songs and the most popular songs. So, we calculate the mean, median, and standard deviation of all songs listed on Spotify and the most popular songs to see how the attributes are shifting with popularity. Also, for better visualization, we plotted the histograms as well. The below table is representing the mean, median, and standard deviation for all listed songs and the most popular songs.

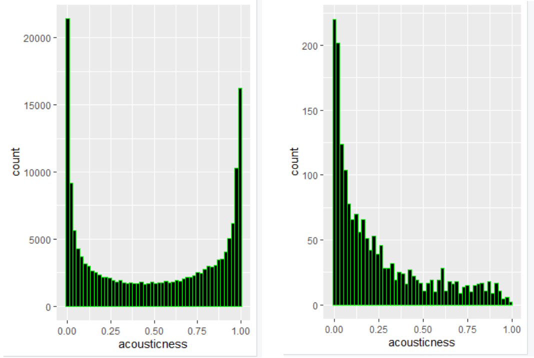
Table for mean, median, and SD values for all song datasets and popular song dataset



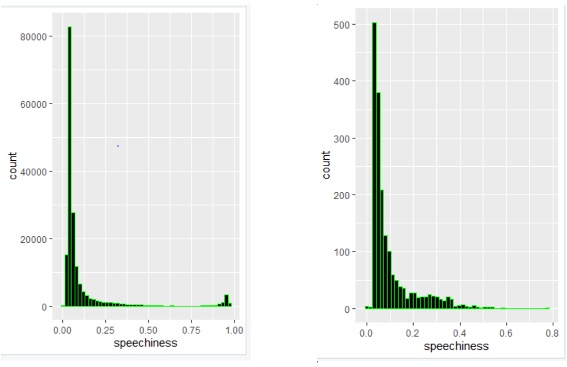
	Parameter	All songs mean	Max popular songs mean	All songs median	Max popular songs median	All songs sd	Max popular songs sd
1	acousticness	0.499	0.248	0.517	0.149	0.38	0.265
2	speechiness	0.106	0.098	0.046	0.057	0.182	0.097
3	duration_ms	232810	210277	205787	205733	148395	49914
4	key	5.205	5.243	5	5	3.518	3.59
5	liveness	0.211	0.17	0.138	0.12	0.18	0.127
6	valence	0.525	0.492	0.536	0.483	0.264	0.23
7	energy	0.483	0.625	0.465	0.64	0.273	0.186
8	danceability	0.537	0.651	0.548	0.665	0.176	0.151
9	instrumentalness	0.197	0.018	0.001	0	0.335	0.099
10	loudness	-11.751	-6.799	-10.836	-6.179	5.692	3.148
11	tempo	117.007	120.354	115.816	119.992	30.254	29.705

Here, we can see that mean values are shifting significantly (either increasing or decreasing) for all songs and most popular songs for all attributes, except some attributes, for which, it is remaining almost the same like speechiness, key, liveness, valence, & tempo. However, a significant difference between the standard deviation says that for most popular songs attribute values are lying in close vicinity of the mean. This can be better visualized from the histograms, as below. (Left image is for all listed songs dataset and the right-side image is for the maximum popular songs dataset)

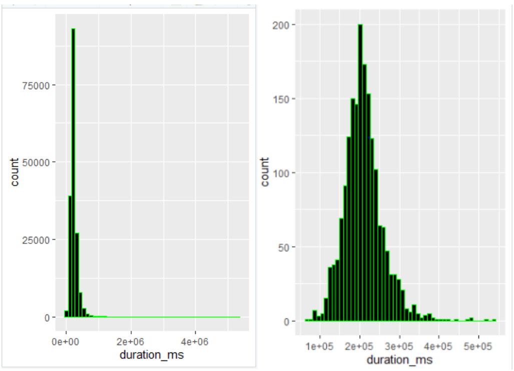
Acousticness



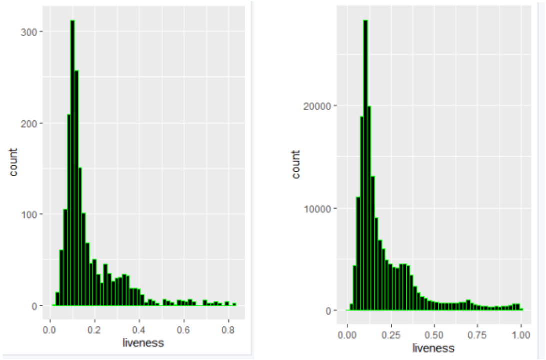
Speechiness



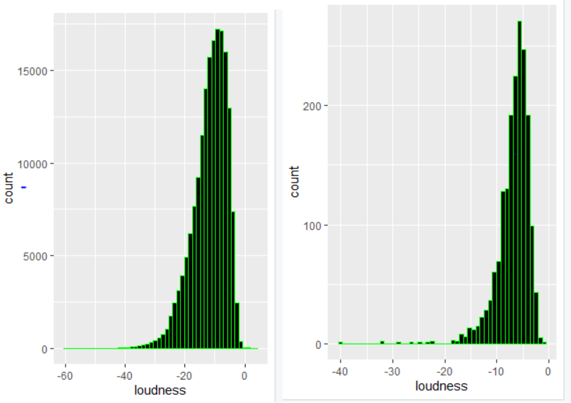
Duration ms



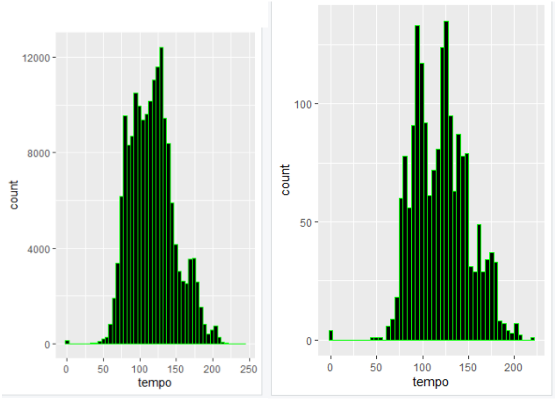
Liveness



Loudness



Tempo



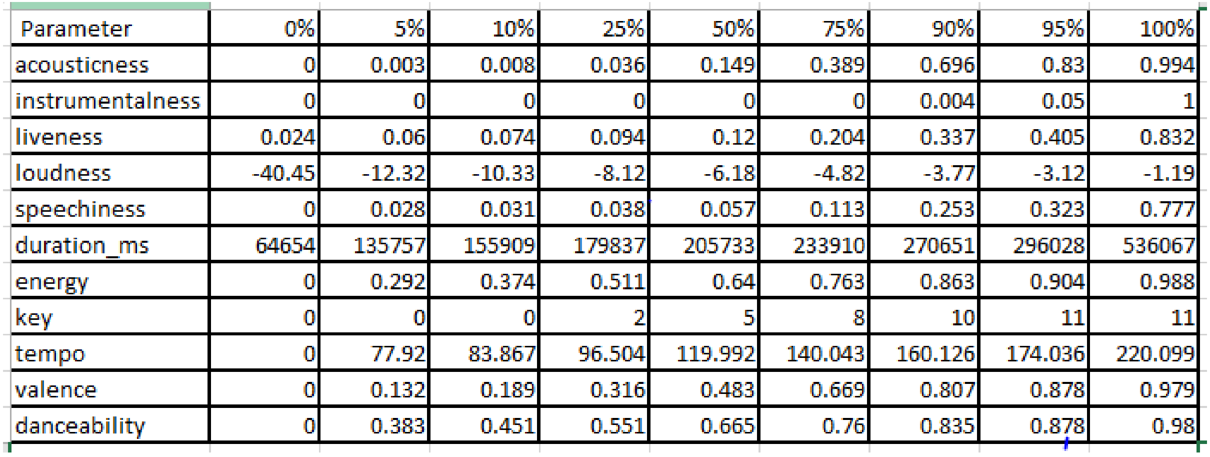
From the table of values and histogram, positive and negative shifts in mean values are as below:

A positive shift in mean values	key, energy, danceability, loudness, tempo
A negative shift in mean values	Acousticness, speechiness, duration_ms, liveness, valence, instrumentalness, danceability

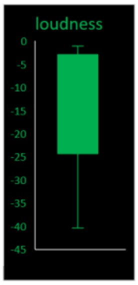
3.3.2. Determining the attribute range for popular songs

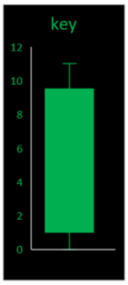
Now, our objective is to determine the range of attribute values in which the maximum popular songs are lying. To determine those values we use the quantile approach to see where the data is lying for 0%, 5%, 10%, 25%, 50%, 75%, 90%, 95%, and 100% data is lying. From histogram plots, we can see that many histograms are either left-skewed or right-skewed, so here median is a better indicator of the central tendency of data, so we used the quantile approach. Also, we plotted box plots for better visualization.

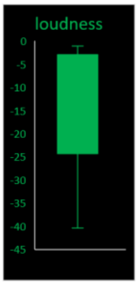
The table and box plots are as below.

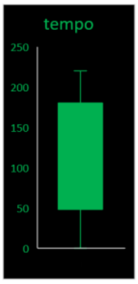


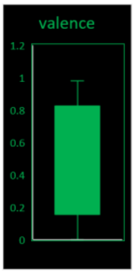
Parameter	0%	5%	10%	25%	50%	75%	90%	95%	100%
acousticness	0	0.003	0.008	0.036	0.149	0.389	0.696	0.83	0.994
instrumentalness	0	0	0	0	0	0	0.004	0.05	1
liveness	0.024	0.06	0.074	0.094	0.12	0.204	0.337	0.405	0.832
loudness	-40.45	-12.32	-10.33	-8.12	-6.18	-4.82	-3.77	-3.12	-1.19
speechiness	0	0.028	0.031	0.038	0.057	0.113	0.253	0.323	0.777
duration_ms	64654	135757	155909	179837	205733	233910	270651	296028	536067
energy	0	0.292	0.374	0.511	0.64	0.763	0.863	0.904	0.988
key	0	0	0	2	5	8	10	11	11
tempo	0	77.92	83.867	96.504	119.992	140.043	160.126	174.036	220.099
valence	0	0.132	0.189	0.316	0.483	0.669	0.807	0.878	0.979
danceability	0	0.383	0.451	0.551	0.665	0.76	0.835	0.878	0.98

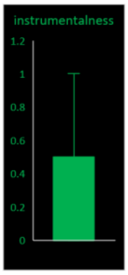


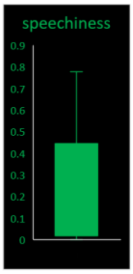


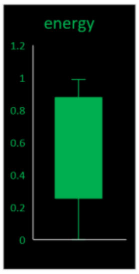


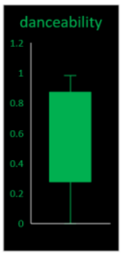


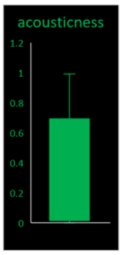


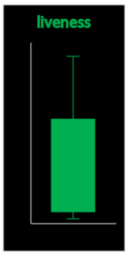




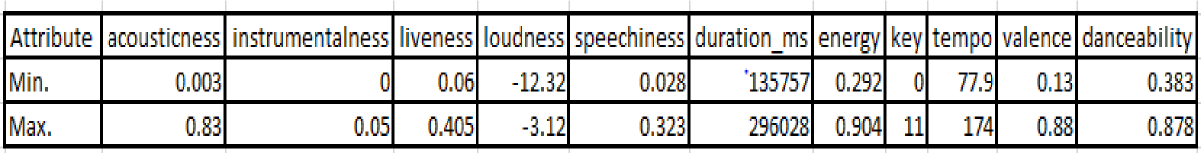








Now, to determine the attribute range for popular songs, first we considered a 90% interval around the median values and calculated the percentage of total popular songs lying in that range. The range about the median is as below,



Attribute	acousticness	instrumentalness	liveness	loudness	speechiness	duration_ms	energy	key	tempo	valence	danceability
Min.	0.003	0	0.06	-12.32	0.028	135757	0.292	0	77.9	0.13	0.383
Max.	0.83	0.05	0.405	-3.12	0.323	296028	0.904	11	174	0.88	0.878

Considering this range, we found that there are **42.84%** of popular songs out of the total popular songs lie in this range.

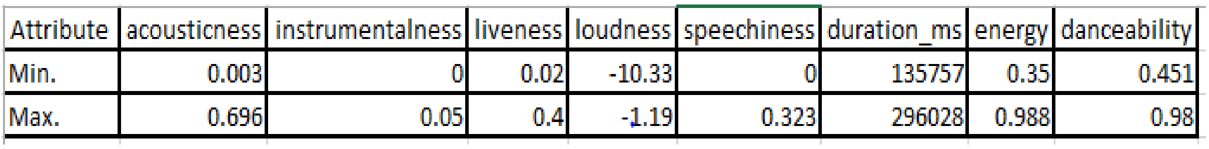
However, for attributes namely key, tempo, and valence there is no significant difference between the entire population mean and most popular song mean, also there is no significant difference between the standard deviation as well for the 2 datasets. So, there will not be any significant difference between the central 90% data minimum and maximum values. This indicates that these attributes are not affecting popularity. So, we exclude those attributes for determining the range.

After excluding these attributes, we got **51.15%** of popular songs to fall in this range.

However, from the heatmap above, we have concluded that there is a positive and negative correlation between the attributes and population. For those having positive correlation we took a 10% to 100% range, for attributes having negative correlation, we considered a 0% to 90% range of values, and for attributes having zero or low correlation values we considered 5% to 95% (central 90% values) for determining the range. After that by hit and trial, we tried to change the range to fit the maximum popular songs in the defined range.

After many iterations, we found the range below, for which, we were able to fit **62.16%** of popular songs, which is a significant number.

The range for attributes is as below,



Attribute	acousticness	instrumentalness	liveness	loudness	speechiness	duration_ms	energy	danceability
Min.	0.003	0	0.02	-10.33	0	135757	0.35	0.451
Max.	0.696	0.05	0.4	-1.19	0.323	296028	0.988	0.98

Conclusion:
In today's environment music industry is becoming more and more competitive. Past data analysis can help composers and producers to understand the current trends. From the dataset, the most of popular songs are released in the past 3 decades. The model explained above helps determine the listener's preference, which will help music composers and producers to produce the song as per the current trend thus making the track probable to lie popular category.

Reference

- https://jovian.ai/zhangxm963/spotify-dataframe/v/39/files?filename=zerotopandas-course-project.ipynb
- https://stackoverflow.com/