UTM
UNIVERSITI TEKNOLOGI MALAYSIA

Project 2
SECI1143 - PROBABILITY AND STATISTICAL DATA ANALYSIS

Semester 2, 2023/2024

Section 01
Group: Ketupat

NAME	MATRIC NUMBER
AMELIA ADLINA BINTI AZRUL	A23CS0043
NAZATUL NADHIRAH BINTI SABTU	A23CS0144
PHAVANEE KATRIYA PHON-AMNUAISUK	A23CS0170
NURUL ATHIRAH SYAFIQAH BINTI MOHD RAZALI	A23CS0163
WAN NUR RAUDHAH BINTI MASZMANIE	A23CS0195

Lecturer: Dr. Sharin Hazlin binti Huspi

Date: 3 July 2024

Link: https://youtu.be/khoSZC3ejnw

Table of Content

1.0 BACKGROUND

The purpose of this study is to analyze the characteristics of the most streamed songs on Spotify in 2023. By examining various attributes of these songs, such as beats per minute (BPM), danceability, acousticness, appearance in Apple charts , and the key they are in, we aim to identify patterns and relationships that might explain their popularity. This analysis could provide insights into what makes a song popular on a major streaming platform like Spotify. We expect to find significant relationships between the different musical attributes and the popularity of the songs, measured by the number of streams. Specifically, we anticipate that the average BPM of popular songs might be higher than a certain threshold. Additionally, there might be a correlation between the danceability and acoustic-ness of songs. We also hypothesize that the appearance in the Apple charts could be a predictor of the number of streams it receives. Finally, we expect that the key of a song might influence its average number of streams.

2.0 DATASET

The dataset that we have chosen for our project is Most Streamed Spotify Songs 2023. This dataset contains a list of the most popular songs as listed by Spotify. It spans 24 columns and has various data about each track, from its name to its musical qualities. A summary of the processed dataset is given in the table below:

Name	Variable	Data type
track_name	The name of the song	Nominal
artist(s)_name	Name of the artist(s) of the song	Nominal
key	Key of the song	Nominal
streams	Total number of streams on Spotify	Ratio
danceability_%	Percentage indicating how suitable the song is for dancing	Ratio
in_apple_charts	Presence and rank of the song on Apple Music charts	Ratio
acousticness_%	Percentage indicating amount of acoustic sound in the song	Ratio
bpm	Beats per minute, a measure of song tempo	Ratio

Selected variables	Test	Description
bpm	Hypothesis testing (one sample test)	**Explanation**: The variable is used to test whether the mean beats per minute of a song is greater than 120 at 5% significance level. **Possible outcome**: The mean beats per minute of a song is greater than 120 at a 5% significance level.
danceability_%, acousticness_%	Correlation test	**Explanation**: The variables are used to test whether there exists a linear relationship between how suitable the song is for dancing and the amount of acoustic sound in the song using Pearson's Product-Moment Correlation at 5% significance level. **Possible outcome**: There is a strong negative linear relationship between how suitable a song is for dancing and the amount of acoustic sound in the song at a 5% significance level. The higher the danceability of the song, the less acoustic it is.
streams, in_apple_charts	Regression test	**Explanation**: The variables are selected to test whether the presence of a song in the Apple Music charts depends on the number of streams on Spotify, using streams as dependent variable Y and in_apple_charts as independent variable X. **Possible outcome**: The amount of Spotify streams is dependent on the presence and popularity of a song on the Apple Music charts. The greater the presence of a song on the Apple Music charts, the more streams it will have on Spotify.

key	ANOVA	**Explanation**: A random sample is selected from the list of songs for different keys to test if the mean value of streams is the same for all different keys. **Possible outcome**: The mean value of streams is the same for all different keys at 5% significance level.
key	ANOVA	**Explanation**: A sample is taken from the top 300 and bottom 300 streams, with the mean streams for each key calculated. This sample is used to test if the mean value of streams in the top 300 and bottom 300 is the same for all different keys. **Possible outcome**: The mean value of streams is different for some keys at 5% significance level.

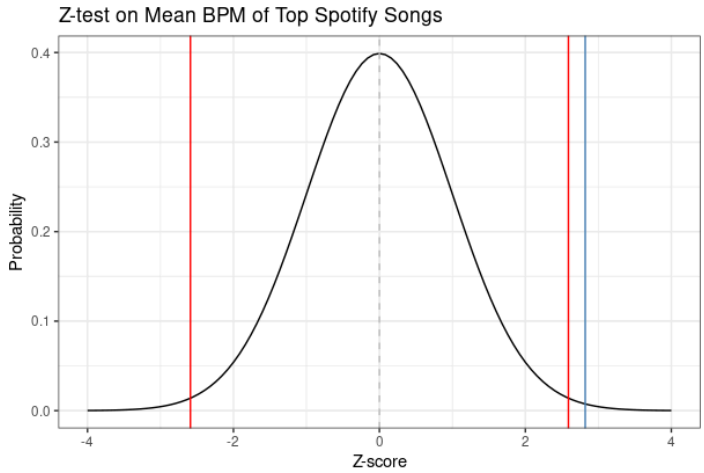
3.0 DATA ANALYSIS

3.1 Hypothesis Testing - One Sample

Based on the test, we wish to determine whether the mean beats per minute (BPM) of a song in the top charts is equal to 120 BPM.

Let μ = mean value of BPM of a song
$H_0 : \mu = 120$
$H_1 : \mu \neq 120$

Significance level, $\alpha = 0.05$
$Z\alpha_{/2} = 2.82$



Graph 1: Normal distribution of Top Spotify Songs

Although the variance of the population is unknown, the sample size, n = 951 is large and we assume that the population is normally distributed by the central limit theorem. Hence, with a normal distribution, we used a Z-test to test the null hypothesis. Using R, we obtained the Z critical values, $Z_{a/2} = 1.96$ and $-Z_{a/2} = -1.96$, indicated by the red vertical lines on the left and right ends of the above graph. Next, we obtain the test statistic for Z, with $Z_0 = 2.82$, indicated by the blue line at the right end of the above graph. Since the test statistic $Z_0 = 2.82 > Z_{a/2} = 1.96$, we reject the null hypothesis. There is enough evidence at a 5% significance level to conclude that the mean value for the BPM of a song is not equal to 120.

3.2 Correlation Test

In this correlation analysis, the variables that we used are danceability and acousticness. We will test whether there is a linear relationship between how suitable the song is for dancing and the amount of acoustic sound in the song using α=0.05 significance level.

Hypothesis statement:

H_0 : $\rho = 0$ (no linear correlation)

H_1 : $\rho \neq 0$ (linear correlation exists)

Test statistics:

By using RStudio, we get the test statistic that shows t is -7.4998. From t-table, since this is two tailed test, there are two critical values:

Lower tail critical value $-t_{\alpha/2} = 0.025, 949 = -1.9625$

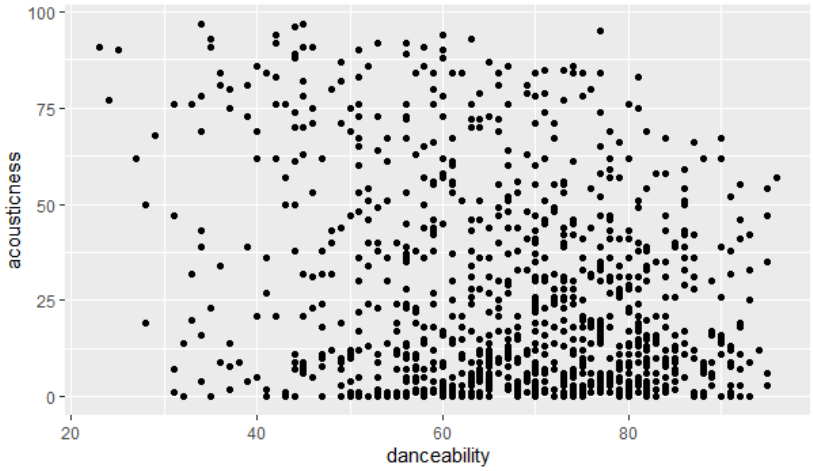
Higher tail critical value $t_{\alpha/2} = 0.025, 949 = 1.9625$

From RStudio we also get p-value = 1.466×10^{-13}

Thus, the H_0 will be rejected if $t < -1.9625$ or $t > 1.9625$. Otherwise, fail to reject H_0 .

Conclusion :

Since $t = -7.4998 < -1.9625$, we reject H_0. There is sufficient evidence to conclude that there is linear relationship between danceability and acousticness at 5% level of significance level.



Graph 2: Scatter plot showing the acousticness by danceability

From the scatter plot, we can see that the point is slope and a lot at the bottom which denotes that there is a negative relationship between danceability and acousticness. The relationship shows that when danceability increases, the acousticness tend to decrease.

Both variables are ratio data type. The sample correlation coefficient then is calculated by using Pearson's correlation coefficient and we got r = -0.2365461 which shows there is relatively weak relationship between the danceability and acousticness.

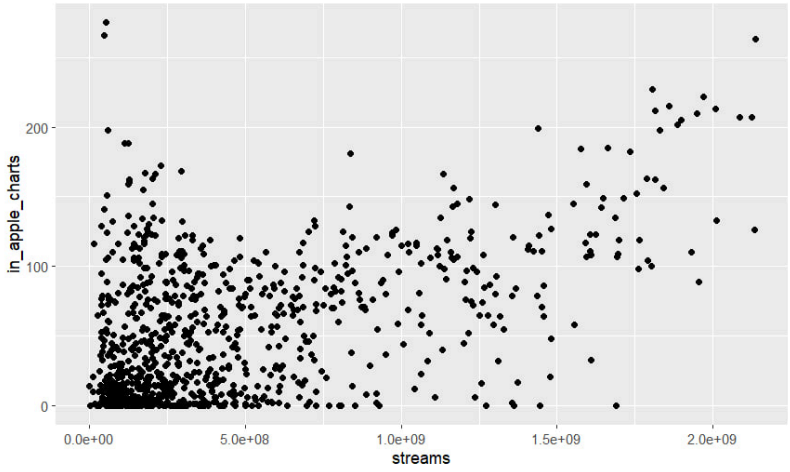
```
        Pearson's product-moment correlation

data:  datase1$danceability_. and datase1$acousticness_.
t = -7.4998, df = 949, p-value = 1.466e-13
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.2956708 -0.1756161
sample estimates:
       cor
-0.2365461
```

3.3 Regression Test

We used the data of Spotify charts to perform a regression test. As we used only a single independent variable, this is a linear regression model. Through this test, we wanted to find out whether there is a linear relationship between the number of streams and the presence in Apple charts. The dependent variable, which is denoted as y, is the number of streams, while the independent variable, x, is the number of appearances in Apple charts.

The following is the scatter plot of presence of a song in Apple charts against the number of streams.



Graph 3 : in Apple charts against streams

Based on the plot, we can see the points are somewhat scattered around the line, indicating that there is a moderate linear relationship between the independent variable (number of appearances in Apple charts) and the dependent variable (number of streams). Through analysis, we obtain the value of the intercept coefficient (b_0) as 220,445,175 and the value of the estimated change in the average number of streams (b_1) as 4,501,330. The estimated regression model is as below:

$$\hat{y} = 220\,445\,175 + 4\,501\,330x$$

From the equation, we estimate that the average number of streams will increase by 4,501,330 for each one-unit increase in the number of appearances in Apple charts. When the number of appearances in Apple charts is zero, the (b_0) value of 220,445,175 indicates that for the songs within the range of streams observed, 220,445,175 is the average number of streams not explained by the number of appearances in Apple charts.

$$R^2 = \frac{\sum(\hat{y} - \bar{y})^2}{\sum(y_i - \bar{y})^2} = 0.2491$$

The coefficient of determinant, R^2 value which indicates that approximately 24.91% of the variation in the number of streams can be explained by the number of appearances in Apple charts, shows a moderate relationship. Therefore,streams on Spotify are reflective of a song's popularity across different streaming platforms, as shown by how a popular presence in Apple charts can moderately predict the number of streams on Spotify.

3.4 ANOVA Test

3.4.1 ANOVA 1

In this data analysis, ANOVA test is used to test the mean value of streams is the same for all different keys from a random sample of songs. The population is split into 11 keys, with NA not included, and the samples are randomly selected to test if all of the mean value is the same for all different keys. The significance level used to test the null hypothesis is $\alpha = 0.05$

The null hypothesis goes as, H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6 = \mu_7 = \mu_8 = \mu_9 = \mu_{10} = \mu_{11}$. While the alternative hypothesis, H_1 :at least one mean is different. The number of samples(n), mean of samples ($\bar{x}$) and standard deviation of samples (s) for each sample are calculated as below.

Sample	n	$\bar{x}$	s
A	74	380403446	376235661
A#	56	442828913	415600864
B	81	424849337	428426977
C#	120	495484198	504908206
D	81	478517818	478912769

D#	33	553036540	562937684
E	62	520207298	535477797
F	89	468446389	471203124
F#	73	461331406	461125101
G	96	411964533	40675000
G#	91	429755352	421082577

Table 3.4.1: The n, x̄ and s for each sample

The mean between samples is $\bar{\bar{x}}$ = 460620475.454545 , the standard deviation between samples is $s_{\bar{x}}$ = 457490148.215667, variance between samples is $ns_{\bar{x}}^2$=159077830891951968.00 , and variance within samples is s_P^2=214834265331430496.00

$$\text{Test statistic:} \quad F = \frac{\text{variance between sample}}{\text{variance within sample}} = \frac{nS_{\bar{X}}^2}{S_P^2}$$

By using the above formula, the value of $F_{\text{test statistic}}$ = 0.7404677. Then the critical value of F with $\alpha = 0.05$ is obtained from the F-distribution table which goes by $F_{\text{critical value}}$ = 1.841891. From the result, since $F_{\text{test statistic}} < F_{\text{critical value}}$ (0.7404677 < 1.841891), we fail to reject the null hypothesis, H_0 as there is sufficient evidence to claim that the list of songs for different keys have the same mean value of streams for all different keys.

3.4.2 ANOVA 2

Previously, ANOVA test was used to test the mean value of streams is the same for all different keys from a random sample of songs. Now, we decided to use the ANOVA test again but this time, the sample is taken from the top 300 and bottom 300 streams, with the mean streams for each key calculated. This sample is used to test if the mean value of streams in the top 300 and bottom 300 is the same for all different keys.

Sample (top 300)	n	x̄	s
A	16	990777077	328377884
A#	20	867772685	380006902
B	25	920733647	407529610
C#	38	1090108794	461053427

D	27	1003031901	479771910
D#	11	1255442450	424986673
E	23	1070375579	486721405
F	31	997694825	422679231
F#	26	942198832	442071306
G	25	993754296	360851362
G#	28	918597397	432178711

Table 3.4.2.1: The n, $\bar{x}$ and s for each sample for top 300 songs

The mean between samples is $\bar{\bar{x}}$ = 1004589771.18182 , the standard deviation between samples for top 300 is $s_{\bar{x}}$ = 429246636.896418, variance between samples is $ns_{\bar{x}}^2$=193561091377256736.00 , and variance within samples is $s_{\square}^2$=179059756853377824.00

By plugging into the formula, the value of $F_{\text{test statistic}}$ = 1.08099. Then the critical value of F with α= 0.05 is obtained from the F-distribution table which goes by $F_{\text{critical value}}$ = 1.86737. From the result, since $F_{\text{test statistic}} < F_{\text{critical value}}$ (1.08099 < 1.86737), we fail to reject the null hypothesis, H_0 as there is sufficient evidence to claim that the list of songs for different keys have the same mean value of streams for all different keys.

Sample (bottom)	n	$\bar{x}$	s
A	26	88596286	38351694
A#	19	90900850	40917836
B	33	92436812	44380379
C#	39	85807690	33449201
D	21	97889399	38471488
D#	9	114736811	31136143
E	21	100218246	40027997
F	30	92038209	41903613

F#	25	100416984	38463399
G	27	89310199	38871760
G#	28	96243155	39480833

Table 3.4.2.2: The n, x̄ and s for each sample for bottom 300 songs

The mean between samples is $\bar{\bar{x}}$ = 95326785.54545 , the standard deviation between samples for top 300 is $s_{\bar{x}}$= 38986139.83065, variance between samples is $ns_{\bar{x}}^2$=1136369940423893.00 , and variance within samples is s□²=1508221809268694.00

By plugging into the formula, the value of $F_{\text{test statistic}} = 0.75345$. Then the critical value of F with $\alpha = 0.05$ is obtained from the F-distribution table which goes by $F_{\text{critical value}} = 1.86627$. From the result, since $F_{\text{test statistic}} < F_{\text{critical value}}$ (0.75345 < 1.86627), we fail to reject the null hypothesis, H_0 as there is sufficient evidence to claim that the list of songs for different keys have the same mean value of streams for all different keys.

4.0 CONCLUSION

Throughout Project 2, we learned that analyzing the characteristics of Spotify's most streamed songs provides valuable insights into what makes a song popular. We applied various statistical techniques, such as hypothesis testing, correlation analysis, regression modeling, and ANOVA tests, to identify patterns and relationships between song attributes and their popularity. This comprehensive analysis improved our understanding of how specific musical elements can influence streaming success on platforms like Spotify.

Our most interesting findings revealed that the mean BPM of popular songs significantly differs from the expected 120 BPM, indicating a trend towards specific tempos. We found a weak negative correlation between danceability and acousticness, suggesting that more danceable songs are less acoustic. Regression analysis showed a moderate relationship between Spotify streams and presence in Apple Music charts, highlighting the impact of cross-platform popularity. Additionally, ANOVA 1 indicated no significant difference in mean streams across different musical keys, while ANOVA 2, comparing top and bottom 300 streamed songs, similarly showed no key-based differences, suggesting that the key does not strongly affect a song's streaming performance.

5.0 APPENDIX

Raw dataset

track_nam	artist(s)_n	artist_cou	released_y	released_r	released_c	in_spotify	in_spotify	streams	in_apple_c	in_apple_c	in_deezer_	in_deezer_	in_shazan	bpm	key	mode	danceabil	valence_%	energy_%	acousticn	instrumer	liveness_%	speechiness_%
Seven (fea	Latto, Jun	2	2023	7	14	553	147	1.4E+08	43	263	45	10	826	125	B	Major	80	89	83	31	0	8	4
LALA	Myke Tow	1	2023	3	23	1474	48	1.3E+08	48	126	58	14	382	92	C#	Major	71	61	74	7	0	10	4
vampire	Olivia Rod	1	2023	6	30	1397	113	1.4E+08	94	207	91	14	949	138	F	Major	51	32	53	17	0	31	6
Cruel Sum	Taylor Swi	1	2019	8	23	7858	100	8E+08	116	207	125	12	548	170	A	Major	55	58	72	11	0	11	15
WHERE SH	Bad Bunny	1	2023	5	18	3133	50	3E+08	84	133	87	15	425	144	A	Minor	65	23	80	14	63	11	6
Sprinter	Dave, Cen	2	2023	6	1	2186	91	1.8E+08	67	213	88	17	946	141	C#	Major	92	66	58	19	0	8	24
Ella Baila S	Eslabon A	2	2023	3	16	3090	50	7.3E+08	34	222	43	13	418	148	F	Minor	67	83	76	48	0	8	3
Columbia	Quevedo	1	2023	7	7	714	43	5.8E+07	25	89	30	13	194	100	F	Major	67	26	71	37	0	11	4
fukumean	Gunna	1	2023	5	15	1096	83	9.5E+07	60	210	48	11	953	130	C#	Minor	85	22	62	12	0	28	9
La Bebe - F	Peso Plun	2	2023	3	17	2953	44	5.5E+08	49	110	66	13	339	170	D	Minor	81	56	48	21	0	8	33
un x100to	Bad Bunny	2	2023	4	17	2876	40	5.1E+08	41	205	54	12	251	83	F#	Minor	57	56	72	23	0	27	5
Super Shy	NewJeans	1	2023	7	7	422	55	5.8E+07	37	202	21	5	168	150	F	Minor	78	52	82	18	0	15	7
Flowers	Miley Cyru	1	2023	1	12	12211	115	1.3E+09	300	215	745	58	1,021	118		Major	71	65	68	6	0	3	7
Daylight	David Kus	1	2023	4	14	3528	98	3.9E+08	80	156	182	24	1,281	130	D	Minor	51	32	43	83	0	9	3
As It Was	Harry Style	1	2022	3	31	23575	130	2.5E+09	403	198	863	46		174	F#	Minor	52	66	73	34	0	31	6
Kill Bill	SZA	1	2022	12	8	8109	77	1.2E+09	183	162	161	12	187	89	G#	Major	64	43	73	5	17	16	4
Cupid - Tw	Fifty Fifty	1	2023	2	24	2942	77	5E+08	91	212	78	6	0	120	B	Minor	78	76	59	43	0	34	3
What Was	Billie Eilish	1	2023	7	13	873	104	3.1E+07	80	227	95	24	1,173	78		Major	44	14	9	96	0	10	3
Classy 10	Feid, Your	2	2023	3	31	2610	40	3.4E+08	43	100	54	14	187	100	B	Major	86	67	66	14	0	12	16
Like Crazy	Jimin	1	2023	3	24	596	68	3.6E+08	8	104	23	2	29	120	G	Major	63	36	73	0	0	36	4
LADY GAG	Gabito Bal	3	2023	6	22	332	26	8.6E+07	11	163	10	4	0	140	F	Minor	65	87	74	22	0	42	4
I Can See '	Taylor Swi	1	2023	7	7	516	38	5.2E+07	73	119	42	1	150	123	F#	Major	69	82	76	6	0	6	3
I Wanna B	Arctic Mor	1	2013	1	1	12859	110	1.3E+09	24	98	582	2	73	135		Minor	48	44	42	12	2	11	3

Table 5.0: Sample of raw dataset used

Processed dataset

track_name	artist(s)_name	streams	in_apple_charts	bpm	key	danceability_%	acousticness_%
Seven (feat. Latto)	Latto, Jung Kook	141381703	263	125	B	80	31
LALA	Myke Towers	133716286	126	92	C#	71	7
vampire	Olivia Rodrigo	140003974	207	138	F	51	17
Cruel Summer	Taylor Swift	800840817	207	170	A	55	11
WHERE SHE GOES	Bad Bunny	303236322	133	144	A	65	14
Sprinter	Dave, Central Cee	183706234	213	141	C#	92	19
Ella Baila Sola	Eslabon Armado,	725980112	222	148	F	67	48
Columbia	Quevedo	58149378	89	100	F	67	37
fukumean	Gunna	95217315	210	130	C#	85	12
La Bebe - Remix	Peso Pluma, Yng	553634067	110	170	D	81	21
un x100to	Bad Bunny, Grupo	505671438	205	83	F#	57	23
Super Shy	NewJeans	58255150	202	150	F	78	18
Flowers	Miley Cyrus	1316855716	215	118		71	6
Daylight	David Kushner	387570742	156	130	D	51	83
As It Was	Harry Styles	2513188493	198	174	F#	52	34
Kill Bill	SZA	1163093654	162	89	G#	64	5
Cupid - Twin Ver.	Fifty Fifty	496795686	212	120	B	78	43
What Was I Made	Billie Eilish	30546883	227	78		44	96
Classy 101	Feid, Young Miko	335222234	100	100	B	86	14
Like Crazy	Jimin	363369738	104	120	G	63	0
LADY GAGA	Gabito Ballestero	86444842	163	140	F	65	22
I Can See You (Tay	Taylor Swift	52135248	119	123	F#	69	6
I Wanna Be Yours	Arctic Monkeys	1297026226	98	135		48	12
Peso Pluma: Bzrp	Bizarrap, Peso Plu	200647221	152	133	F	85	26
Popular (with Play	The Weeknd, Mad	115364561	182	99	C#	85	7
SABOR FRESA	Fuerza Regida	78300654	149	130	G	79	9
Calm Down (with	Rï¿½ï¿½ma, Selen	899183384	119	107	B	80	43
MOJABI GHOST	Tainy, Bad Bunny	61245289	109	122	F#	81	14
Last Night	Morgan Wallen	429829812	107	204	F#	52	46
Dance The Night (	Dua Lipa	127408954	0	110	B	67	2

Table 5.1: Sample of processed dataset of the amount of streams, track name in apple chart, bpm, key, danceability and acousticness