Giacomo De Gobbi Matricola: 860913
Davide Vercesi Matricola: 852483

PROGETTO FOUNDATION OF PROBABILITY AND STATISTICS

ANALISI DISCOGRAFIA RED HOT CHILI PEPPERS SU SPOTIFY

ABSTRACT

Con questo progetto abbiamo studiato e analizzato alcune caratteristiche tecniche e non di tutta la discografia della band californiana. Per fare ciò è stata utilizzata la libreria di R spotifyr che ha permesso di collegarsi alla API della famosa applicazione Spotify, ricavando svariate informazioni riguardanti gli 11 album e le 182 canzoni del gruppo. Abbiamo cercato di ricavare pattern e correlazioni tra queste variabili ed infine si è indagato sulla presenza di una possibile relazione tra la popolarità dei brani e tali variabili.

La band

I Red Hot Chili Peppers (talvolta abbreviato semplicemente in RHCP o Red Hot) sono un gruppo rock statunitense, formatosi a Los Angeles nel 1983, attualmente composto da Anthony Kiedis (voce), Flea (basso, cori), John Frusciante (chitarra, cori) e Chad Smith (batteria, percussioni). Nella loro carriera hanno mescolato con successo vari generi, tra cui soprattutto funk, rap, hard rock, punk rock e successivamente alternative rock e pop rock, arrivando a forgiare un caratteristico sound che nelle esibizioni dal vivo è improntato spesso all'improvvisazione. Hanno venduto più di 80 milioni di dischi nel mondo e nel 2012 sono stati inseriti nella Rock and Roll Hall of Fame.

Spotifyr

Spotifyr è una libreria R dalla quale si possono estrarre in blocco le funzionalità audio delle canzoni e altre informazioni dall'API Web di Spotify. Raggruppando automaticamente le richieste API, consente di inserire il nome di un artista e di recuperare la sua intera discografia in pochi secondi, insieme alle caratteristiche audio di Spotify e alle metriche di popolarità del brano / album. Permette inoltre di estrarre informazioni su brani e playlist per un determinato utente Spotify. (1)

In questo lavoro sono stati utilizzati due comandi della libreria:

- get_audio_features (2) con cui sono state estratte le caratteristiche musicali della discografia

- get_tracks_popularity con cui è stato estratto l'indice di popolarità delle singole canzoni

I risultati del primo comando sono stati salvati in un file .csv e successivamente si è aggiunta la colonna riguardante la popolarità delle canzoni.
Abbiamo anche provveduto al ad eliminare alcune colonne/informazioni da questo file perché irrilevanti per il lavoro (es. codici id e url degli album e delle canzoni di Spotify).
Il dataset finale consiste perciò di 182 osservazioni (le canzoni della band) e 16 variabili.

---

[1] Per maggiori informazioni: https://github.com/charlie86/spotifyr

[2] Per maggiori informazioni: https://developer.spotify.com/documentation/webapi/reference/tracks/get-audio-features/

Le variabili

NOME VARIABILI	TIPO VARIBILI	DESCRIZONE VARIABILI
Album_name		Titolo dell'album
Album_release date		Anno di uscita dell'album
Track_name		Titolo della canzone
danceability	Quantitativa continua (0-1)	Descrive quanto sia adatta una canzone ad essere ballata su una combinazione di elementi musicale come il tempo, ritmo, forza del beat e regolarità. Un valore di 0.0 è il meno ballabile e 1.0 è il più ballabile
energy	Quantitativa continua (0-1)	Rappresenta una misura percettiva di intensità e attività. Tipicamente, canzoni energetiche risultano veloci, potenti e rumorose. Altre caratteristiche percettive che contribuiscono a questo attributo sono il range dinamico, il timbro e l'entropia generale.
loudness	Quantitativa continua (-60-0)	Indica la rumorosità/potenza media della canzone in decibels. I valori tipicamente variano in un range da -60 a 0 decibels
speechness	Quantitativa continua (0-1)	Individua la presenza di parti parlate nella canzone. Più la canzone assomiglia ad un discorso più il valore sarà vicino a 1.0. Valori sopra 0.66 descrivono canzoni che sono probabilmente composte solo da parti parlate. Valori tra 0.33 e 0.66 indicano che la canzone sembra contenere sia parte musicale che parte parlata. Valori al di sotto di 0.33 indicano la sola presenza di sole parti musicali

acousticness	Quantitativa continua (0-1)	Indica il grado di parti acustiche nella canzone. Se il valore è prossimo a 1.0 c'è un'altissima probabilità che la canzone sia acustica
instrumentalness	Quantitativa continua (0-1)	Indica se la canzone contenga o no parti vocali. Più il valore è vicino a 1.0, maggiore è la probabilità che la canzone non contenga contenuti vocali.
liveness	Quantitativa continua (0-1)	Individua la presenza di un pubblico nella registrazione. Alti valori della variabile indicano un aumento di probabilità che la canzone è stata eseguita live. Un valore superiore a 0.8 fornisce un'alta probabilità che la canzone è live.
valence	Quantitativa continua (0-1)	Misurazione da 0.0 a 1.0 che descrive la positività espressa da una canzone. Canzoni con alti valori di valence suonano più positive mentre canzoni con valori bassi suonano più negative
tempo	Quantitativa continua	La stima generale del ritmo di una canzona misurata in battiti per minuto (BPM). Nella terminologia musicale il tempo è la velocità o il ritmo di una data parte e deriva direttamente dalla durata media del beat.
Duration ms	Quantitativa continua	La durata della canzone in millisecondi
Key_mode	Qualitativa ordinale	Indica la chiave tonica della canzone
Track_popularity	Quantitativa discreta (0-100)	Indica la popolarità della canzone, il valore varia da 0 a 100. La popolarità è calcolata da un algoritmo di Spotify ed è basata sul numero totale di riproduzioni e su quanto queste siano recenti

Studio chiave tonica

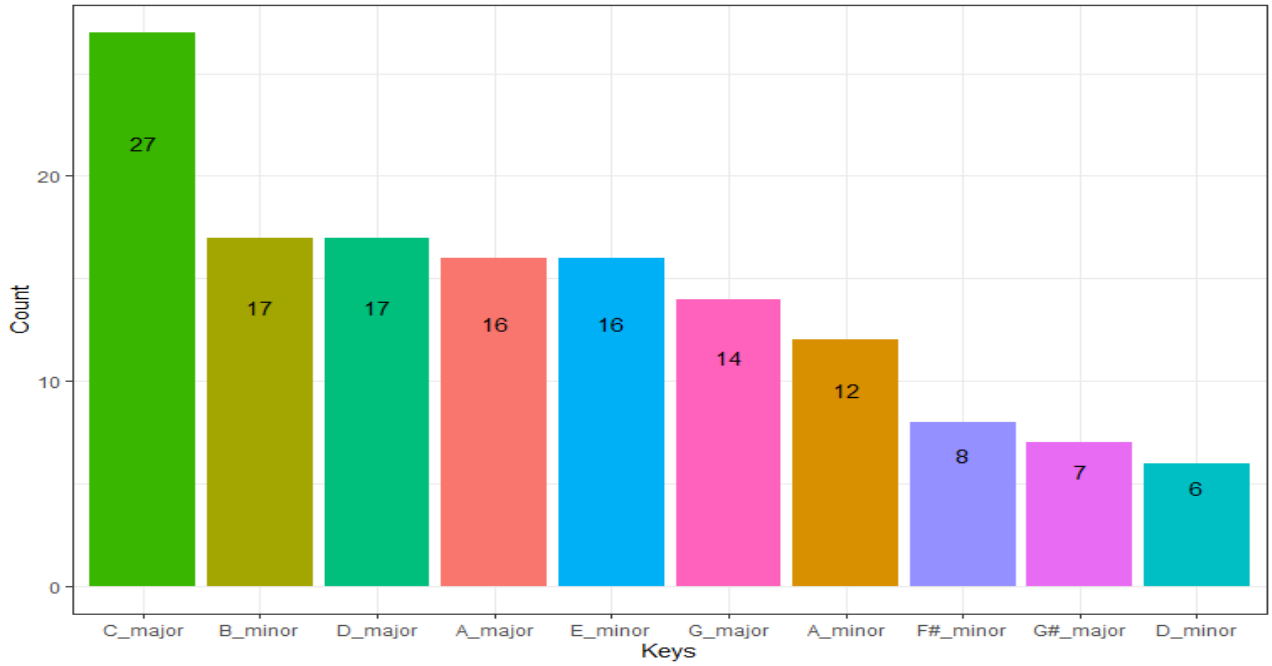


Figure 1: Barchart

Dal grafico a barre notiamo che sono state usate 10 differenti chiavi, 5 maggiori e 5 minori. La più utilizzata è C major (Sol maggiore) la meno utilizzata D minor (Re minore). Ciò non sorprende dato che la chiave in Sol maggiore è la più utilizzata nella musica moderna perché più adatta a strumenti come il pianoforte e la chitarra. Oltretutto, la chiave in Sol maggiore è una delle chiavi più semplici da utilizzare per comporre canzoni e soprattutto, nel rock, si utilizza questo espediente in modo da concentrarsi maggiormente su altri aspetti della canzone come melodia e testo. Questa analisi è corroborata dal lavoro del data analyst Kenny Ning[3], il quale ha analizzato 30 milioni di tracce su Spotify. Un altro aspetto rilevante è che gli accordi maggiori sono notoriamente associati ad atmosfere più solari e serene a differenza degli accordi minori più usati delle canzoni tristi.

Dal nostro istogramma si evince quale sia l'esperienza musicale che la band vuole proporre, cioè che la discografia dei RHCP è divisa in due filoni: quello delle canzoni più allegre ed energiche (Dani California, Give it Away …) e quello delle canzoni più tristi e malinconiche (Under the Bridge, Otherside …)

[3] Per maggiori informazioni: https://www.hypebot.com/hypebot/2015/05/the-most-popular-keys-of-all-music-on-spotify.html

Media e deviazione standard

MEDIA	danceability	energy	loudness	speechiness	acousticness	instrumentalness	liveness	valence	tempo	duration_ms	track_popularity
DISCOGRAFIA	0.5432	0.8246	-5.455	0.085	0.0801	0.0720	0.189	0.499	117.118	245221	48.56
THE GATEWAY (2016)	0.5395	0.7308	-6.069	0.070	0.1763	0.0300	0.176	0.551	119.573	248260	60.92
IM WITH YOU (2011)	0.5610	0.8624	-3.188	0.0562	0.0515	0.0010	0.1878	0.4992	120.094	254534	52.92
STADIUM ARCADIUM (2006)	0.5261	0.7801	-4.526	0.0871	0.1180	0.0049	0.1743	0.5005	119.379	313905	56.62
BY THE WAY (2002)	0.5141	0.8242	-3.934	0.0627	0.0382	0.0009	0.176	0.3849	122.271	257364	58.56
CALIFORNICATION (199)	0.4570	0.833	-3.148	0.1045	0.0632	0.0787	0.1472	0.4626	116.592	225933	60.20
ONE HOT MINUTE (1995)	0.4796	0.7847	-5.588	0.0709	0.0940	0.0441	0.2078	0.3978	113.359	283863	45.07
BLOOD SUGAR SEX MAGIC (1991)	0.5782	0.7495	-11.381	0.0590	0.0071	0.1194	0.1857	0.6335	120.849	260913	55.70
MOTHER MILK (1989)	0.4676	0.9309	-4.171	0.1174	0.0317	0.1621	0.2767	0.4329	124.285	246078	40.78
UPLIFT MOFO PARTY PLAN (1987)	0.6013	0.9155	-4.021	0.1262	0.0155	0.0004	0.1728	0.356	111.339	192096	36.66
FREAKY STYLEY (1985)	0.6327	0.8505	-6.573	0.0762	0.1301	0.1746	0.2160	0.6531	106.650	187875	31.33
RHCP (1984)	0.631	0.8293	-7.240	0.1039	0.1310	0.1579	0.1481	0.5562	110.447	172459	32

DEV. STANDARD	danceability	energy	loudness	speechiness	acousticness	instrumentalness	liveness	valence	tempo	duration_ms	track_popularity
DISCOGRAFIA	0.1439	0.1805	3.318	0.0840	0.1716	0.1984	0.1377	0.2296	26.598	140041	13.86
THE GATEWAY (2016)	0.1962	0.1325	1.053	0.0572	0.2099	0.0598	0.1026	0.2656	32.512	46739	6.44
IM WITH YOU (2011)	0.1264	0.1041	1.208	0.0314	0.0817	0.0037	0.0959	0.1832	22.545	41453	9.00
STADIUM ARCADIUM (2006)	0.1276	0.2134	2.861	0.1712	0.2103	0.0250	0.1079	0.1999	27.321	280975	8.02
BY THE WAY (2002)	0.0939	0.1254	0.996	0.0546	0.0505	0.0019	0.1339	0.1846	22.221	47354	9.72
CALIFORNICATION (199)	0.1325	0.2136	3.026	0.0884	0.1273	0.1620	0.1071	0.2316	22.540	54656	11.32
ONE HOT MINUTE (1995)	0.142	0.2447	2.964	0.0316	0.2571	0.0997	0.1409	0.1964	27.658	80833	9.14
BLOOD SUGAR SEX MAGIC (1991)	0.0971	0.2077	1.259	0.0292	0.0142	0.2361	0.0969	0.2255	30.534	812926	10.40
MOTHER MILK (1989)	0.1239	0.1069	1.519	0.0500	0.0930	0.3286	0.2157	0.2103	30.284	139886	9.00
UPLIFT MOFO PARTY PLAN (1987)	0.0853	0.0909	1.303	0.0643	0.0208	0.0010	0.1035	0.1330	21.109	53204	6.42
FREAKY STYLEY (1985)	0.2054	0.2067	4.542	0.0326	0.2331	0.2913	0.1966	0.2634	31.549	114616	8.41
RHCP (1984)	0.0968	0.1479	2.581	0.0422	0.2384	0.3078	0.1027	0.2420	20.262	71136	10.12

Analizziamo ora i risultati più interessanti trovati utilizzando la media e la deviazione standard dei valori delle singole canzoni divise per gli 11 album.
Innanzitutto, il valore medio della variabile energy (0.8) riguardante l'intera discografia è in linea con il genere e le caratteristiche del gruppo.
Se andiamo a guardare i singoli album notiamo che tutti hanno un valore alto (il minimo è 0.73). Questa informazione, se collegata con la precedente sulle chiavi toniche, ci dice che la discografia è divisa tra canzoni più allegre e brani più tristi, in tutti gli album sono presenti entrambe le tipologie di mood rendendoli bilanciati. Tale affermazione è validata se guardiamo anche la variabile valence. Quest'ultima variabile, che è una sorta di indicatore di positività della canzone, assume in quasi tutti gli album valori vicini a 0.5, la deviazione standard è anche in questo caso per quasi tutti gli album vicina a 0.2, ciò indica un ampio spettro di temi, sonorità e sensazioni nella discografia del gruppo californiano.
Un altro risultato che ci aspettavamo è che il valore medio di speechness è estremamente basso, dato che valori alti di tale variabile si riscontrano nel genere rap e non nel rock. Anche acousticness presenta valori bassi, dato che la band ha preferito nel corso dell'intera carriera utilizzare sonorità più graffianti ed incisive. Riguardo a danceability si nota in media un valore medio di 0.5 e non è casuale che i primi tre album dei Red Hot Chili Peppers abbiano valori superiori a 0.6 dato che sono stati prodotti durante la fase funk/rythmic rock della band, con presenza di sonorità più ballabili e vicine al mondo black. La variabile tempo si aggira attorno ai 120 bpm che è il ritmo in cui si assesta la grande maggioranza della musica rock.
La variabile liveness presenta una media bassa per tutti gli album, il che significa che le canzoni sono state registrate in studio e non durante i concerti.
La variabile che invece sorprende è quella relativa alla popolarità delle canzoni. Infatti, se escludiamo l'album Californication del 1999, notiamo che gli album più recenti contengono canzoni più popolari, mentre gli album più vecchi sono meno popolari. Ciò è in contraddizione, oltre che al nostro gusto personale, con la critica musicale che considera gli ultimi due album i peggiori della discografia della band e assolutamente non paragonabili con il quintetto uscito tra il 1991 e il 2002.
Una possibile ragione di questa differenza potrebbe essere data dal fatto che solo gli ultimi due album sono usciti quando era già presente il servizio Spotify, perciò i fan con questa opzione di riproduzione musicale avrebbero preferito lo streaming al posto della copia fisica, mentre ipoteticamente gli album precedenti erano già presenti nella collezione degli appassionati perciò hanno meno riproduzioni e sono meno popolari.
Un'altra spiegazione potrebbe essere legata al fatto che la riproduzione illimitata di musica di Spotify potrebbe aver avvicinato dei curiosi ad ascoltare, anche se non

appassionati, delle canzoni dei Red Hot e dato che la visualizzazione delle pagine premia gli album più recenti questo potrebbe spiegare la distorsione.
Andiamo ora a visualizzare le informazioni delle precedenti tabelle con l'ausilio di alcuni grafici.

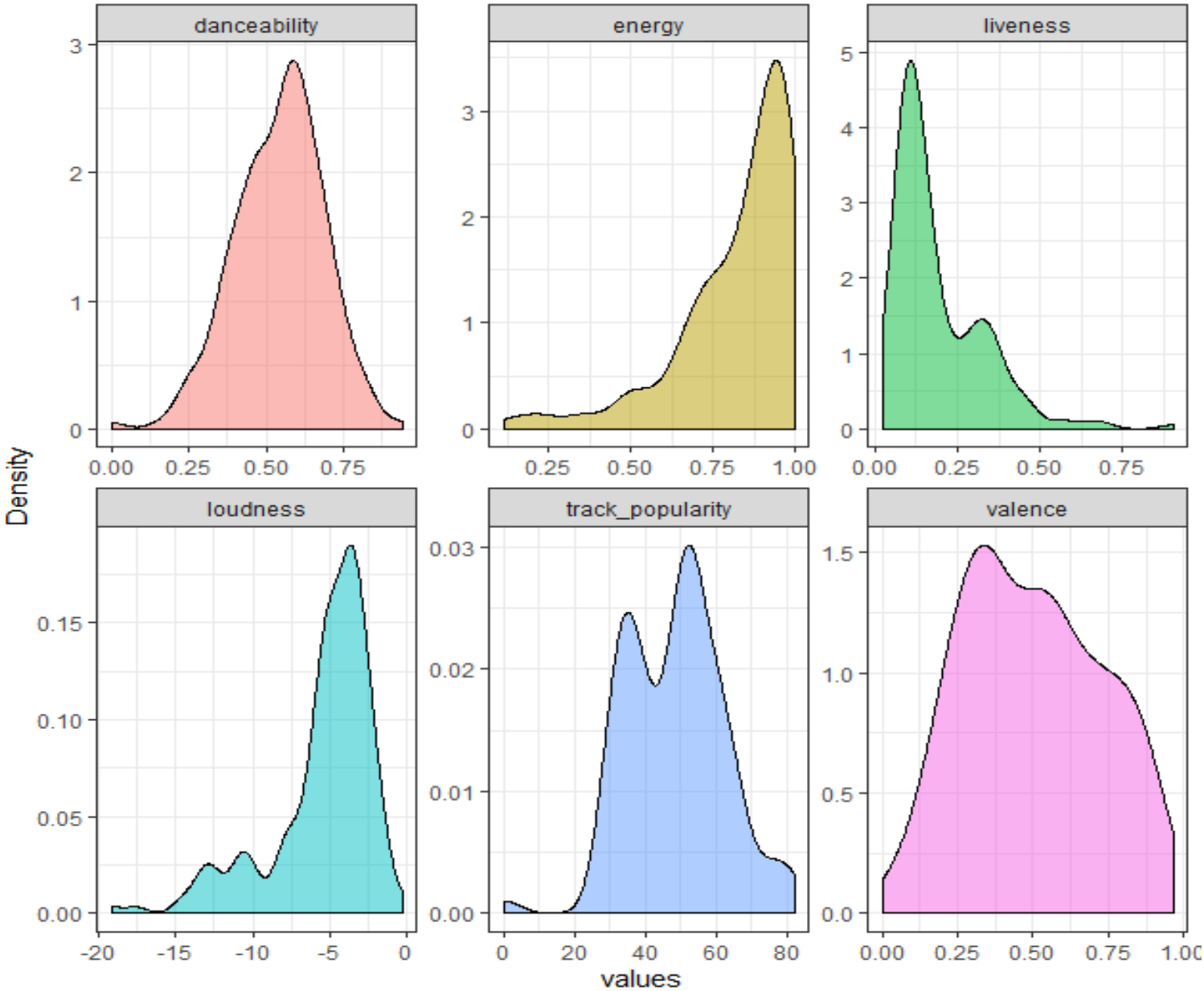


Figure 2: Density plot

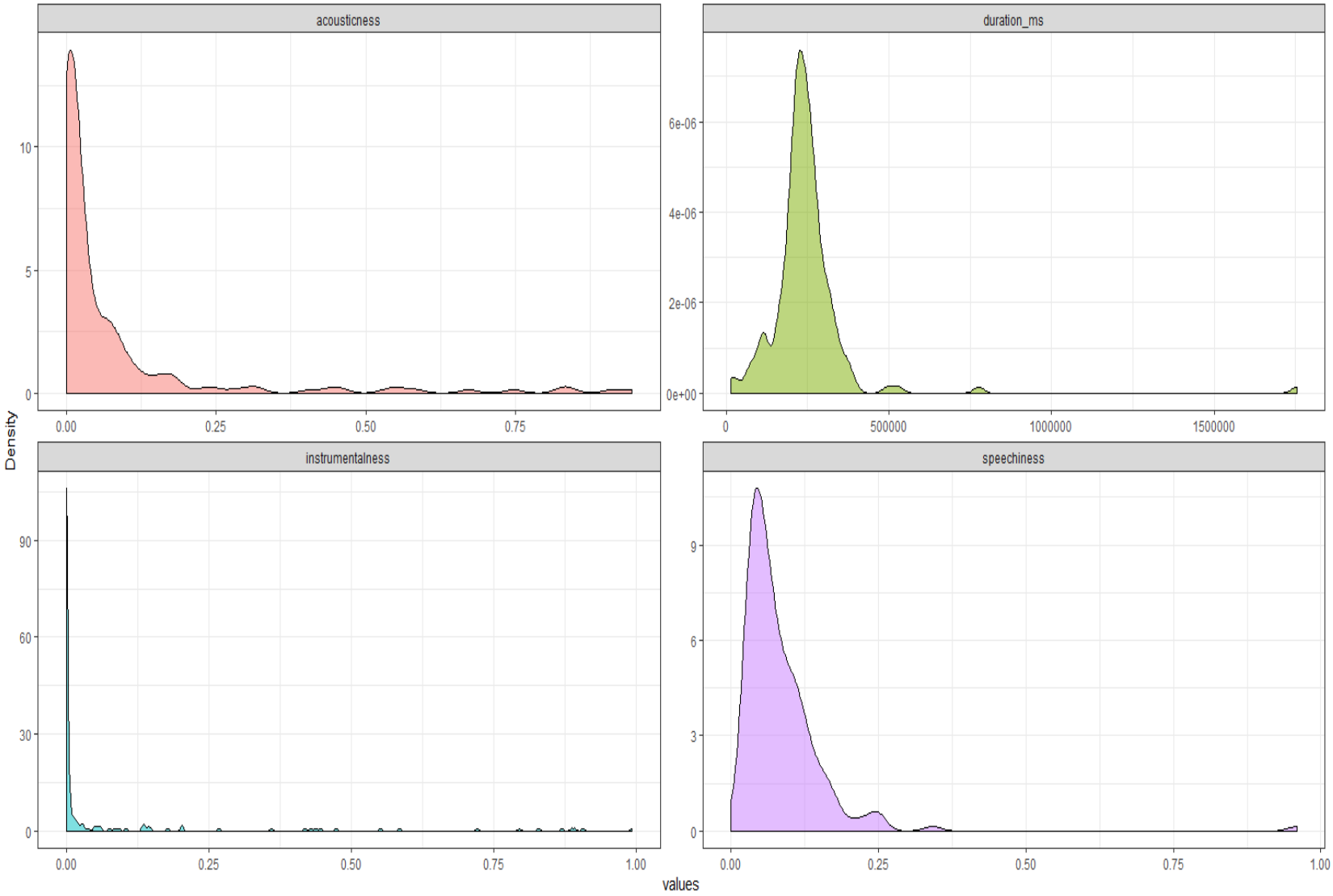


Figure 3: Density plot

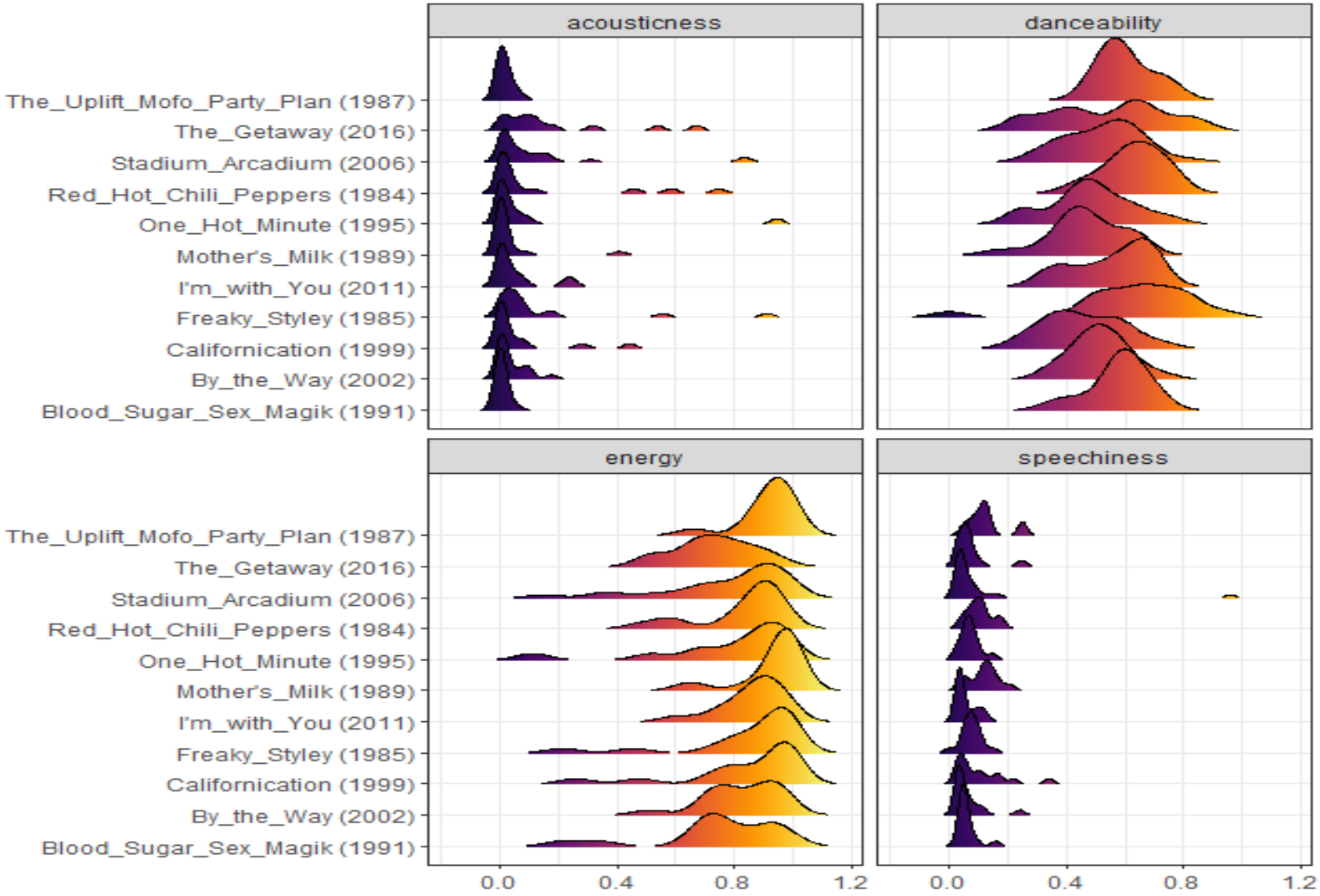


Figure 4: Ridgeline plot

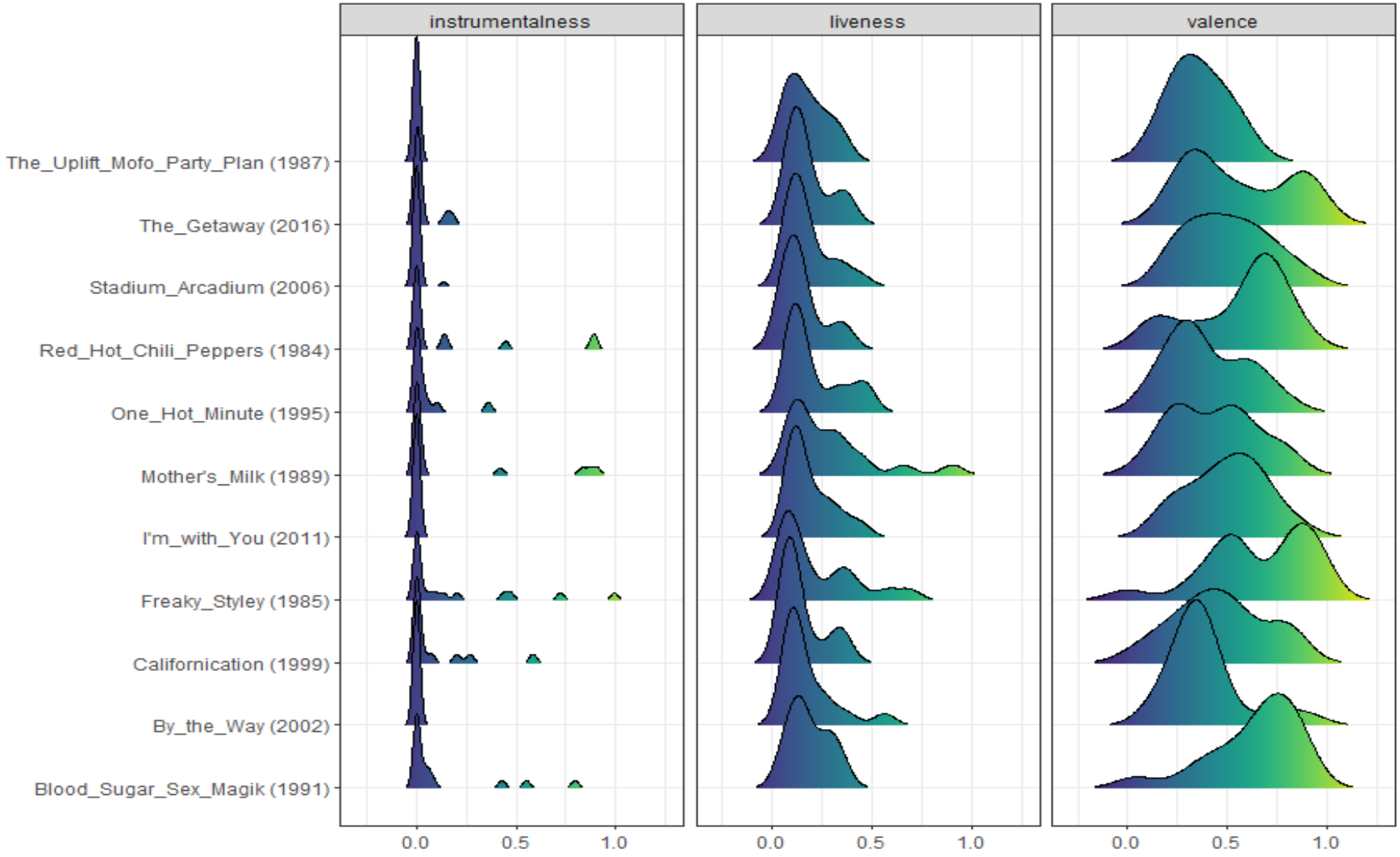


Figure 5: Ridgeline plot

Dai grafici si evince che sono presenti degli outliers, soprattutto per quanto riguarda le variabili acousticness e instrumentalness, la maggior parte si trovano nei primi album 1984-85-89-91, durante la fase funk. Questo è dovuto al fatto che inizialmente la band componeva più assoli con i vari strumenti senza utilizzare la voce del cantante. Per quanto riguarda le altre variabili mantengono una curvatura intorno alla media senza avere degli outliers significativi.

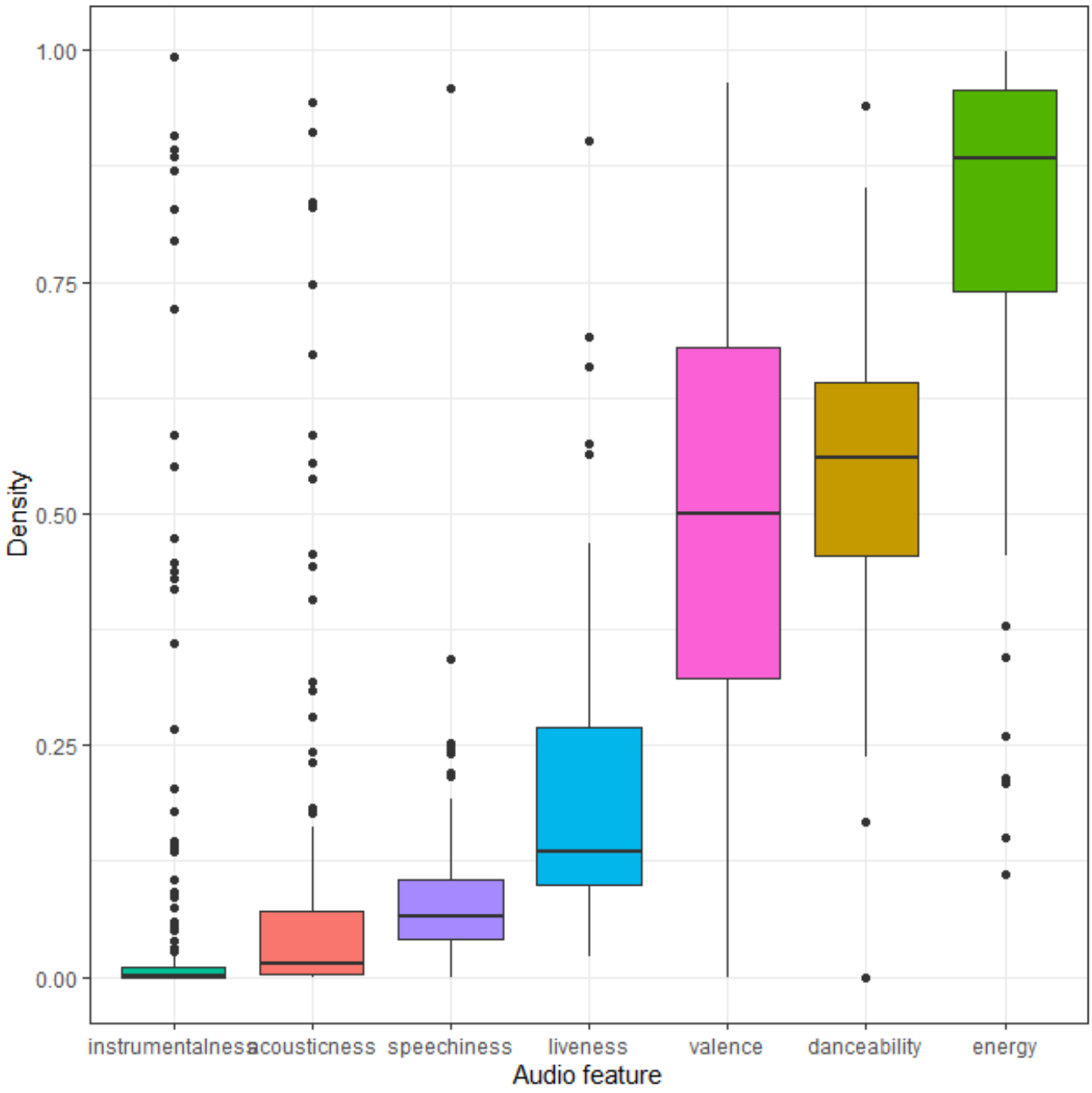
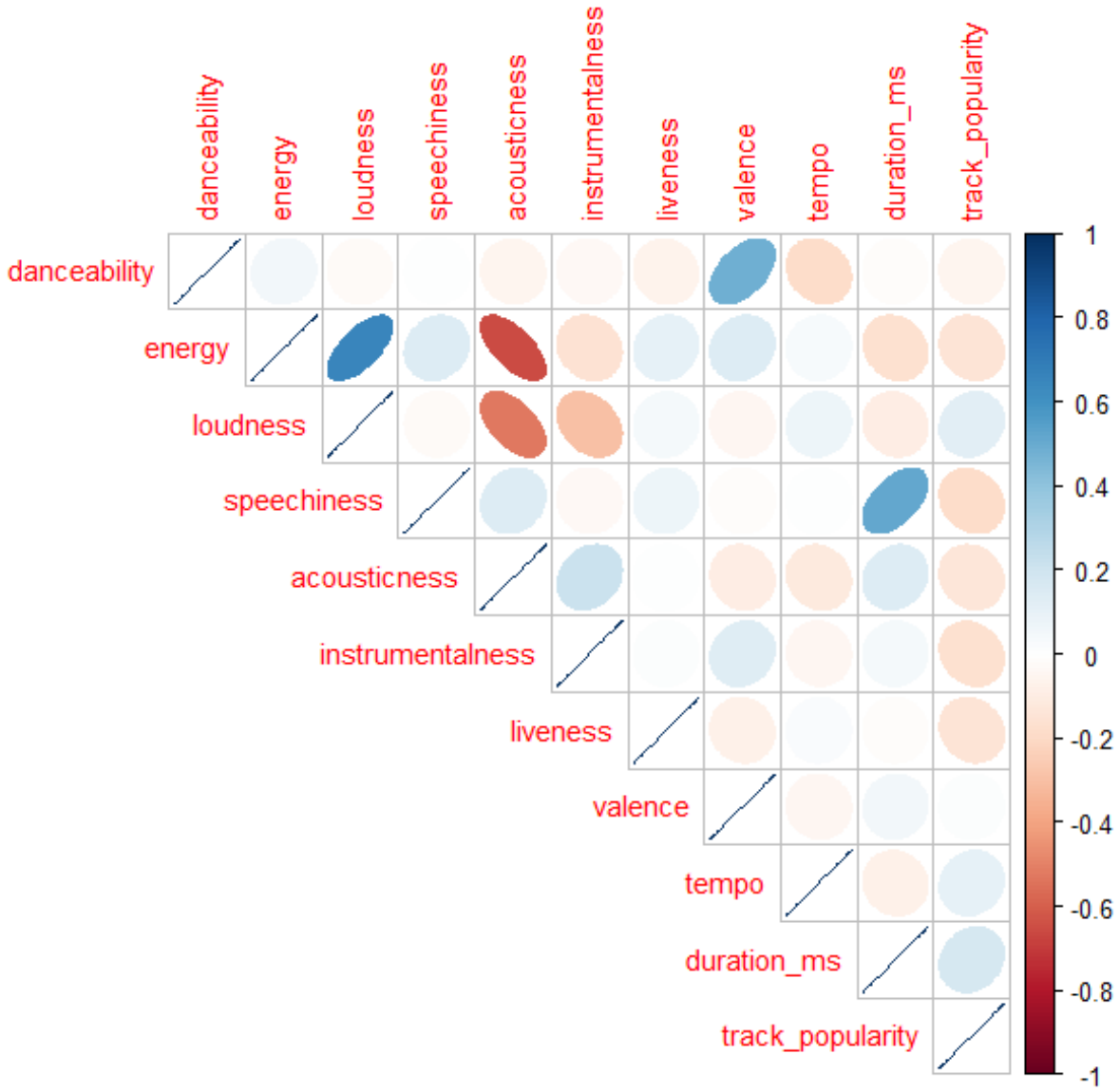


Figure 6: Boxplot

Correlazione e Regressione Lineare

Analizziamo le correlazioni tra le variabili, interessandoci soprattutto sull'eventuale esistenza di correlazioni con la popolarità delle singole canzoni



Come possiamo facilmente notare abbiamo riscontrato delle correlazioni piuttosto triviali. Infatti, è normale che la variabile energy sia correlata positivamente con la variabile loudness e negativamente con acousticness e non sorprende nemmeno il segno positivo tra danceability e valence, dato che per logica più una canzone è

positiva più questa è ballabile. Purtroppo, la variabile track_popularity non sembra avere forti correlazioni con nessuna altra variabile e perciò è difficile stabilire l'esistenza di un pattern che spieghi la popolarità di una canzone attraverso le sue caratteristiche audio/strumentali.
Tentiamo ora una regressione multipla lineare e vediamone i risultati:

ANALISI VARIANZA					
	gdl	*SQ*	*MQ*	*F*	*Significatività F*
Regressione	10	9128,13591	912,813591	6,080546966	6,82573E-08
Residuo	171	25670,57288	150,1203092		
Totale	181	34798,70879			

F-test è significativo perciò almeno una variabile indipendente influenza significativamente la nostra variabile dipendente (track_popularity).
Purtroppo, R quadro corretto presente un valore di 0.22 perciò la quota di devianza (varianza) della variabile dipendente Y spiegata dalla relazione lineare con le variabili esplicative è molto bassa.

	Coefficienti	*Errore standard*	*Stat t*	*Valore di significatività*
Intercetta	76,03137129	11,21530356	6,779252196	1,879E-10
danceability	-7,290486998	7,492175292	-0,973080142	0,331887942
energy	-31,53950474	9,006006347	-3,502052245	0,000589091
loudness	1,102626389	0,396499157	2,780904752	0,006028855
speechiness	-37,60679868	14,50425422	-2,592811605	0,010343848
acousticness	-18,26758276	7,722025284	-2,365646587	0,019118706
instrumentalness	-9,25282432	4,965392563	-1,863462798	0,064111886
liveness	-8,355605073	6,724766009	-1,242512388	0,215748761
valence	6,169652011	4,814174699	1,281559643	0,201732187
tempo	0,046612959	0,035552965	1,311085005	0,191586916
duration_ms	2,81219E-05	8,23515E-06	3,41486625	0,00079703

Infatti, ben 6 variabili risultano non significative.
Come ci si poteva aspettare, data la forte correlazione tra alcune variabili indipendenti, si è presentato il fenomeno della multicollinerità imperfetta (basti osservare gli elevati standard error delle variabili).
Proviamo ora a risolvere tale problema togliendo le variabili energy, loudness e duration dal modello di regressione e vediamo cosa succede.

ANALISI VARIANZA					
	gdl	*SQ*	*MQ*	*F*	*Significatività F*
Regressione	7	3496,374409	499,482058	2,776466352	0,009231057
Residuo	174	31302,33438	179,898473		
Totale	181	34798,70879			

Anche qui F-test è significativo ma, il punteggio del R quadro corretto è ancora più basso (0.06)

	Coefficienti	*Errore standard*	*Stat t*	*Valore di significatività*
Intercetta	52,03318759	6,857142775	7,58817328	1,88288E-12
danceability	-8,804611153	8,153494879	-1,0798573	0,281700258
speechiness	-27,82713549	12,05528343	-2,3082938	0,022159137
acousticness	-4,711039197	6,124016368	-0,7692728	0,442774374
instrumentalness	-11,78760763	5,275041225	-2,2346001	0,026716552
liveness	-12,91199311	7,283911111	-1,7726731	0,078032632
valence	4,29512423	5,103886389	0,84153994	0,401200601
tempo	0,044520885	0,038542298	1,15511757	0,249626102

Stavolta solo due variabili risultano significative e gli errori standard rimangono alti.

Purtroppo, da tale analisi non è si è riusciti a ricavare una relazione lineare tra la popolarità delle canzoni e le variabili messe a disposizione dalla Api di Spotify.
La mancanza di alcune variabili importanti potrebbe essere una risposta per i nostri risultati, ma non sono da escludere le considerazioni fatte in precedenza riguardanti la popolarità degli album. Infatti, la pagina della band su Spotify mette a disposizione in primo piano le cinque canzoni più popolari e questo potrebbe generare un ascolto maggiore, da parte dei “curiosi”, indotto proprio da questa impostazione grafica. Le altre tracce invece devono essere selezionate e cercate manualmente da parte dell’utente, che come già detto, spesso non è un fan o un esperto del gruppo. Se consideriamo anche il fatto che nell’applicazione vengono visualizzati prima gli album più recenti possiamo così chiudere il cerchio ed accettare i risultati trovati e concludere che una analisi concernente l’utilizzo della tecnica di regressione non è saggio utilizzando solamente i dati che ci concede Spotify.
Probabilmente questo effetto sul nostro studio, che è intrinseco alla UXD (user experience design) dell’applicazione, è esacerbato dal fatto che i RHCP abbiamo una carriera che si protrae da oltre 35 anni e che ormai non siano più “mainstream” e di moda come un tempo. Sarebbe perciò estremamente interessante in futuro ripetere questo progetto utilizzando una band o un performer anagraficamente ed artisticamente più giovane e studiarne i risvolti.