La popolarità dei brani di Spotify

Matricole:

Marco Branciforti: **796670**
Emanuele Marnati: **812503**
Valerio Schips: **872954**

Abstract

All'alba del 2021 è possibile affermare che il mercato della musica sia incredibilmente cresciuto ed abbia visto l'entrata di nuovi soggetti concorrenti alle tradizionali società operanti nella distribuzione, prima di tutto grazie ad una diffusione massiva di internet e dei dispositivi tecnologici onnipresenti come gli smartphone. Il mercato della distribuzione musicale del ventunesimo secolo che, in un primo momento, è stato contraddistinto dalla presenza di servizi freeware come "E Mule" o "Lime Wire" è, ad oggi, egemonizzato da servizi a pagamento come Spotify. Allo scopo di comprendere quali caratteristiche influenzano la popolarità di un brano, questo lavoro parte dall'analisi di dati inerenti a più di 160.000 brani presenti su Spotify e prodotti negli ultimi 100 anni. Ipotizzando che il successo di una canzone sia influenzato da caratteristiche tecniche, la seguente ricerca volgerà quindi alla definizione delle determinanti della popolarità del brano.

Introduzione

Spotify ha un servizio gratuito di web API (Application Program Interface) accessibile agli sviluppatori gratuitamente che, tramite la tecnica di scraping integrato, supporta la possibilità di estrarre facilmente dati sui brani, artisti, playlist, livelli di ascolto e variabili più tecniche riferite ai brani quali: dimensioni e lunghezza della canzone, chiavi musicali presenti e molte altre. È stato utilizzato un database ottenuto da Spotify Web API che preleva brani in modo casuale. Proprio per la varietà di generi musicali e per la disponibilità di informazioni presenti su Spotify potrebbe essere un buon dataset sul quale effettuare le nostre analisi e provare a ottenere un modello lineare adatto a spiegare il fenomeno della popolarità dei brani negli ultimi 100 anni. Questo potrebbe permetterci di comprendere quali siano le mosse vincenti comuni agli artisti di successo, siano essi russi, giapponesi o americani o di comprendere la loro difficoltà nell' avere successo.

Dataset

Per ottenere i risultati è stato analizzato il dataset (Community Data Licence Agreement) disponibile sulla piattaforma Kaggle relativo ai brani presentati dall'anno 1921 al 2020, per un totale di 169.909 unità descritte e 18 variabili numeriche, categoriche e binarie. Per ognuna di queste sono state analizzate la distribuzione e alcuni indici.

Variabile	Descrizione	Tipologia
Acousticness	Brano digitale o analogico	Quantitativa Continua
Artists	Nome dell'artista	Qualitativa Nominale
Danceability	Propensione al ballo	Quantitativa Continua
Duration_ms	Durata del brano	Quantitativa Continua
Energy	Intensità della canzone	Quantitativa Continua
Explicit	Presenza di testo esplicito	Qualitativa Dicotomica
Instrumentalness	Preponderanza del cantato sugli strumenti	Quantitativa Continua
Key	Semitono del brano	Qualitativa Ordinale
Liveness	Indica se il brano è in live	Quantitativa Continua
Loudness	Volume in dB	Quantitativa Continua
Mode	Scala melodica utilizzata	Qualitativa Dicotomica
Name	Nome del brano	Qualitativa Nominale
Popularity	Popolarità del brano	Quantitativa Discreta
Release_date	Data di rilascio della canzone	Qualitativa Ordinale
Speechiness	Preponderanza del cantato sui suoni	Quantitativa Continua
Tempo	BPM del brano	Quantitativa Continua
Valence	Positività del brano	Quantitativa Continua
Year	Anno di rilascio	Quantitativa Discreta

Dopo una prima analisi del dataset si è scelto di focalizzarsi sulle canzoni uscite dal 2000 al 2018. Questa scelta è stata adottata per analizzare gli ascolti in un periodo in cui la diffusione della musica avviene con mezzi omogenei, escludendo quindi le canzoni prodotte precedentemente all'avvento di internet come strumento di diffusione di massa. Si sono inoltre esclusi gli anni 2019 e 2020 in quanto i dati presenti erano incoerenti e incompleti. Si è ottenuto quindi un dataset di 34.401 osservazioni.

Sono state escluse le variabili **Name, Artists, Release_Date** perchè sono caratteristiche proprie di ogni singolo brano e quindi risultano essere poco utili ai fini dell'analisi.

Successivamente sono state rimosse le osservazioni frutto di errore nell'estrazione dei dati:

- **Tempo:** la variabile ha un per definizione un range che può essere compreso tra i 60 e i 300 BPM, sono stati quindi esclusi i valori non compresi nell'intervallo.
- **Speechiness:** sono stati esclusi i brani con valore maggiore di 0,6 perché in quasi tutti i casi risultano essere dei podcast.

Dalla descrizione riportata su Kaggle, veniva fatto presente che uno stesso brano poteva essere ripetuto più volte a causa di un remake effettuato negli anni successivi. A questo proposito, è stato fatto un controllo ed effettivamente risultavano brani con stesso autore e stesso titolo ma con caratteristiche differenti. In più, per la gran parte dei casi duplicati, uno di essi ha popolarità pari a zero. Ipotizzando che faccia fede la prima pubblicazione di una canzone, sono state ordinate le tracce per anno di pubblicazione crescente ed eliminati i duplicati degli anni successivi.

Sono state quindi estratte dal dataset con un campionamento casuale semplice il 20% dei record, ottenendo così 6826 osservazioni che saranno oggetto della nostra analisi.

Descrizione e Analisi Variabili

Nel seguente capitolo vengono descritte le variabili, dalle quali sono stati rimossi gli outliers (ad eccezione di popularity che sarà la variabile target del modello) al fine di ottimizzare il modello finale. Sono stati utilizzati entrambi i metodi dello z-score e dell'Inter Quantile Range per la rimozione degli outliers.

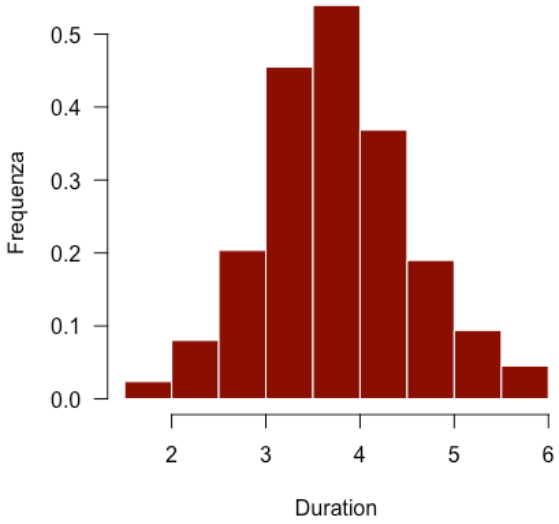
Duration_ms

La variabile duration_ms rappresenta la lunghezza dei brani in millisecondi.

Si è proceduto all'eliminazione degli outliers.

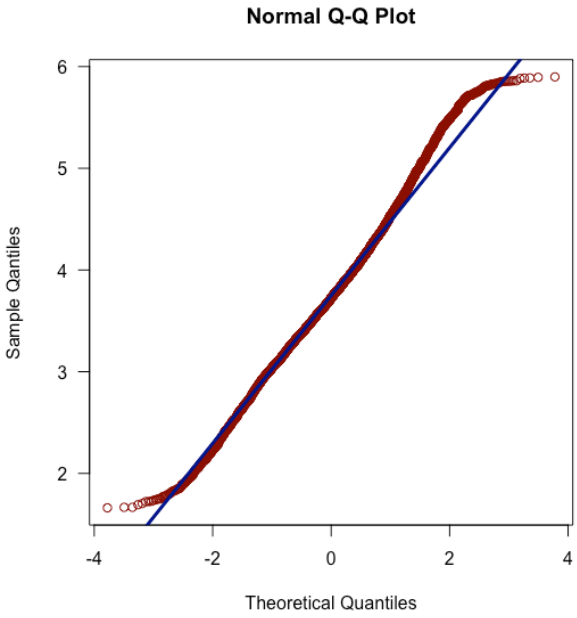
Summary	Duration (min:sec) Con Outliers	Duration (min:sec)
Min.	00:31	01:39
1st Qu.	03:14	03:15
Median	03:44	03:42
Mean	03:51	03:45
3rd Qu.	04:18	04:14
Max	23:13	05:54

Una volta eliminati gli outliers il nostro dataset contiene 6513 osservazioni e la distribuzione delle frequenze di duration_ms risulta la seguente.



Si nota come media e mediana coincidano quasi perfettamente.

Il qqplot evidenzia come nonostante la distribuzione non sia perfettamente simmetrica, la distribuzione si avvicini molto alla normale.



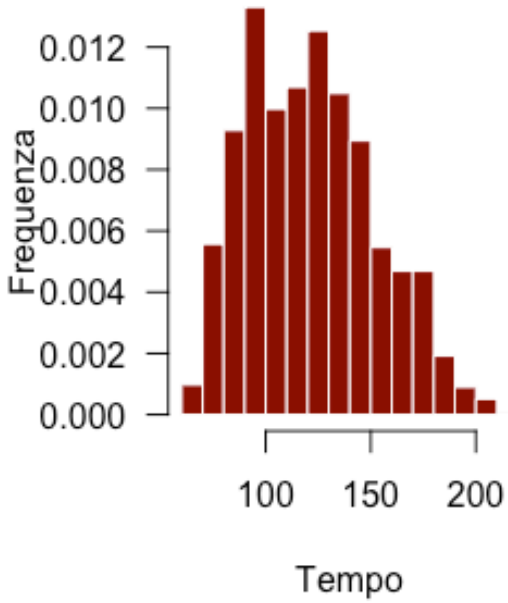
Tempo

Questa variabile indica la frequenza di battiti per ogni minuto (BPM) della traccia. Il range usuale va dai 60 BPM per la musica Reggae ai 300 BPM per un brano Heavy Metal.

Summary	Tempo(BPM)
Min.	60,01
1st Qu.	96,29
Mode	120
Median	120,01
Mean	121,83
3rd Qu.	141,97
Max	210,17

Si nota quindi come la distribuzione delle frequenze sia asimmetrica a destra.

$$Mode < Median < Mean$$

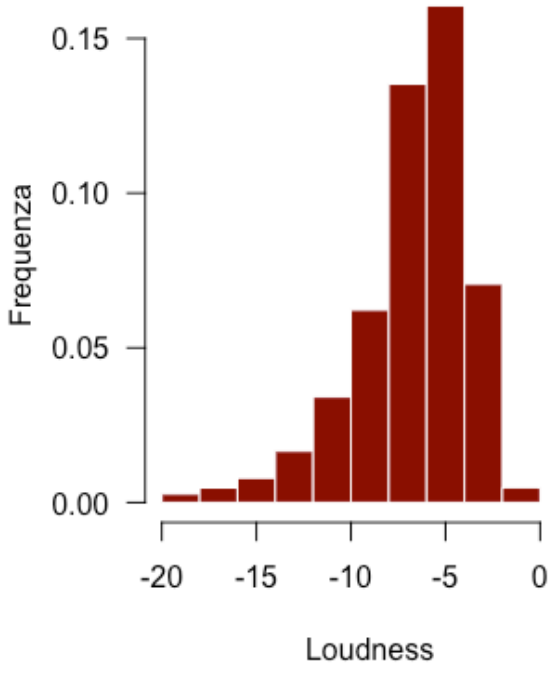


Loudness

Rappresenta una media per ogni traccia dei volumi misurati in decibel. Più il valore è vicino agli 0dB più il volume sarà alto. Il range usuale per questo valore è compreso tra i -60dB e gli 0dB quindi i nostri valori sono coerenti con quanto atteso.

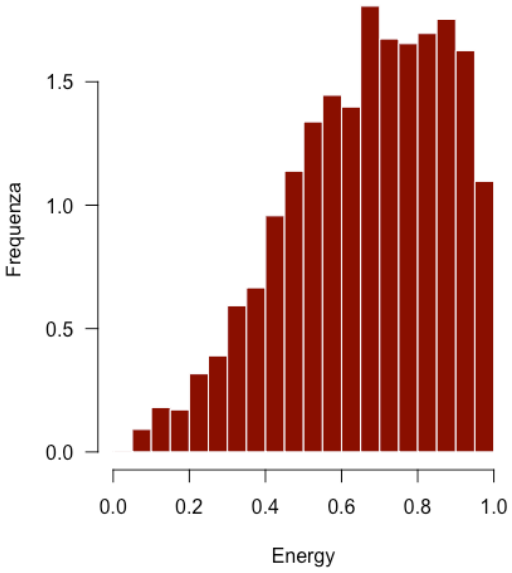
Summary	Loudness (dB) Con Outliers	Loudness (dB)
Min.	-44,281	-19,795
1st Qu.	-8,318	-8,097
Median	-6,243	-6,170
Mean	-7,179	-6,735
3rd Qu.	-4,670	-4,626
Max	0,197	0,197
Mode	-5.167	

Una volta rimossi gli outliers si può notare che la distribuzione delle frequenze è fortemente asimmetrica a sinistra.



Energy

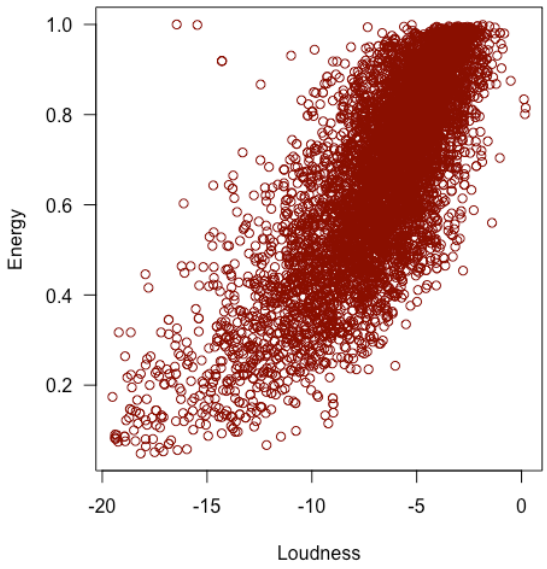
Rappresenta in un range da 0 a 1 l'intensità della canzone. Tipicamente più la traccia è veloce e rumorosa più avremo valori vicino a 1. Ad esempio un brano Death Metal avrà un alto valore in energy mentre un brano classico avrà un valore più vicino al minimo.



Dopo aver rimosso gli outliers il grafico della distribuzione mostra un asimmetria verso sinistra.

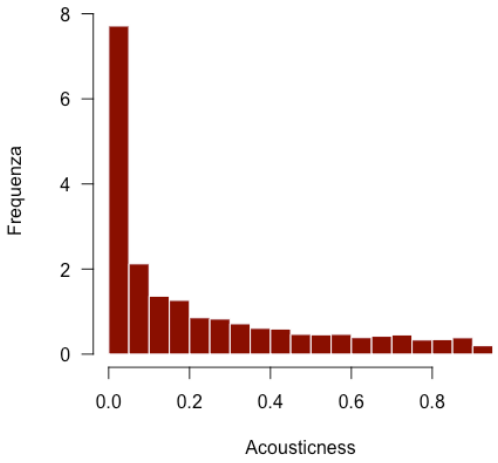
Mean	Median	Mode
0,6639	0,6870	0,877

Ci si aspetta quindi una forte presenza di brani intensi e molto rumorosi, potrebbe essere interessante fare uno scatter plot delle variabili Energy e Loudness per notare se ci sia qualche interazione. A valori bassi di Energy corrispondono valori bassi di Loudness e lo stesso discorso si può fare con i valori più alti. Dal grafico congiunto si potrebbe sospettare un legame di tipo lineare.



Acousticness

Variabile quantitativa che varia tra 0 ed 1 che rappresenta l'utilizzo di soli mezzi acustici vocali o strumentali come ad esempio chitarre classiche, contrapposte all'utilizzo di mezzi elettronici come sintetizzatori ed auto-tuned. Valori bassi indicando brani artificialmente creati, mentre valori alti indicano brani creati da strumenti analogici o vocali naturali.



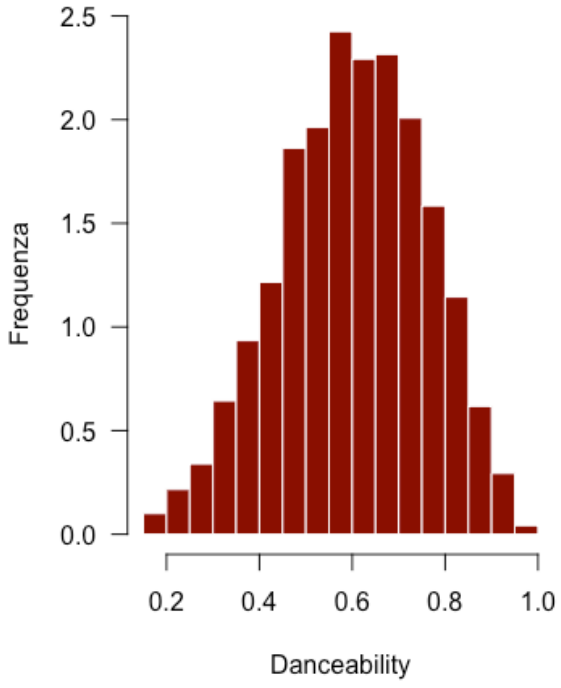
Mean	Median
0,22215	0,106

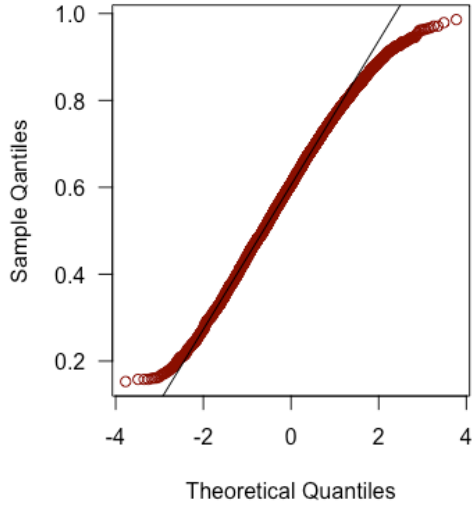
Danceability

Descrive quanto la traccia sia ballabile in un range tra 0 e 1. Gli elementi musicali che la caratterizzano includono il tempo, la stabilità del ritmo, la forza del battito e la regolarità complessiva della traccia.

Dall'analisi della variabile si riscontra una distribuzione che presenta una lieve asimmetria a sinistra, ma la rimozione degli outliers apporta comunque un miglioramento alla distribuzione senza rimuovere un significativo numero di valori.

Mean	Median
0,602	0,607





Come si può notare dall'istogramma e dal qqplot la distribuzione delle frequenze risulta quasi simmetrica, inoltre una parte consistente delle osservazioni sono distribuite sulla retta che identifica la distribuzione normale.

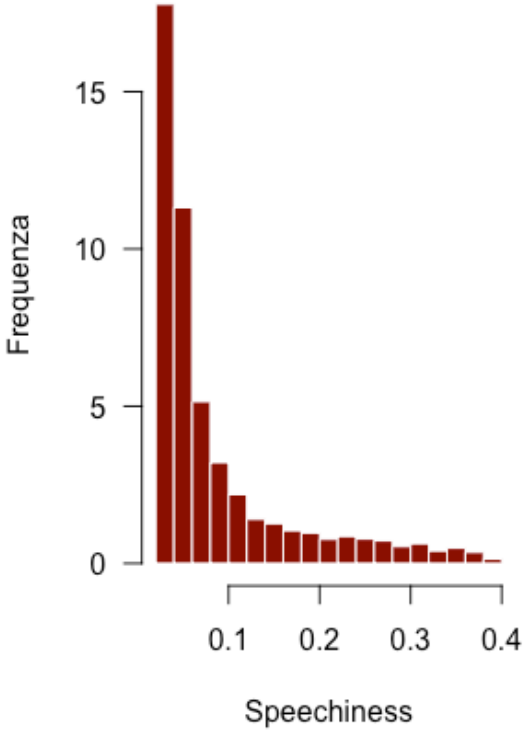
Speechiness

Variabile quantitativa che varia tra 0 e 1 e rappresenta l'utilizzo di parole e voce contrapposto all'utilizzo di suoni. Valori prossimi a uno rappresentano tracce con l'assenza di musica, mentre valori prossimi a 0 figurano musica strumentale concertistica.

La variabile era già stata manipolata nel capitolo precedente, la rimozione di outliers ha ridotto ulteriormente il range della variabile.

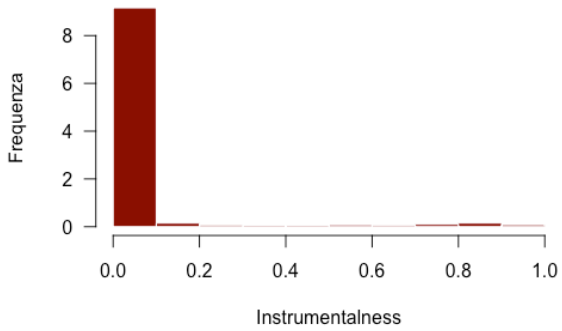
Summary	Speechiness Con Outliers	Speechiness
Min.	0,0228	0,0228
1st Qu.	0,0353	0,0352
Median	0,0519	0,0508
Mean	0,0948	0,0867

3rd Qu.	0,107	0,101
Max	0,597	0,384



Instrumentalness

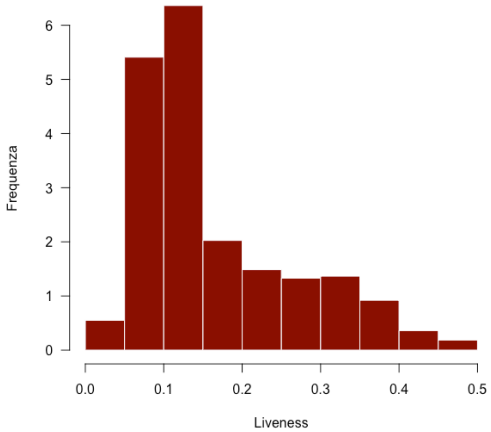
Indica, in un range tra 0 e 1, quanto è alta la presenza del cantato nella traccia. È importante osservare che i suoni come “Ooh” o “Aah” non sono considerati come cantato. Più il valore si avvicina ad 1 maggiore è la preponderanza degli strumenti sul cantato.



Per questa variabile il 90% delle osservazioni cadono nel range tra 0 e 0,05 di cui più del 48% delle ultime ha valore 0. Si è quindi concluso che instrumentalness non poteva fornire nessuna informazione utile all’analisi ed è stata quindi rimossa.

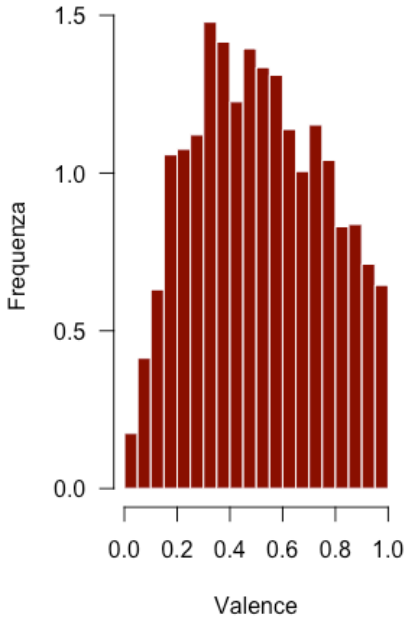
Liveness

Questa variabile rileva la presenza di un pubblico durante la registrazione. Più è probabile che la traccia sia registrata in un concerto, più vicino a 1 sarà il valore dell'attributo. Un valore superiore a 0,8 rappresenta una traccia sicuramente live, mentre valori sotto a 0,6 indicano brani certamente prodotti in studio ma che possono comprendere tracce pre-registrate di un pubblico o la presenza di più voci che si sovrappongono.



Valence

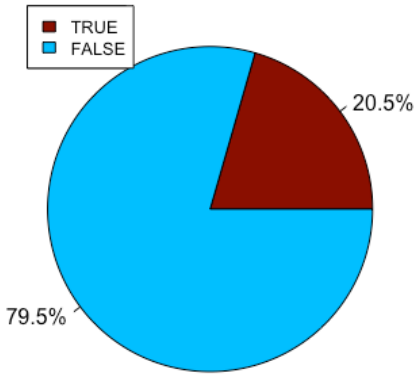
Variabile che descrive la positività musicale trasmessa da una traccia. I brani con valenza alta sono più positivi (felici, allegri, euforici), mentre i brani con valenza bassa suonano musica più negativa (tristi, depressi, arrabbiati).



Questo attributo in combinazione con l'energia è un forte indicatore dell'umore acustico e delle qualità emotive generali che possono caratterizzare l'acustica del brano.

Explicit

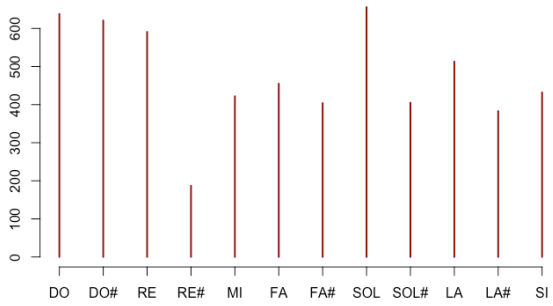
Indica se la traccia abbia o meno testi espliciti.



Key

Variabile categorica e tecnica, rappresenta i 12 valori partizionati, (semitoni) di una ottava (da 0 a 11). La scala si suddivide in 12 gradi progressivi, calcolati ripartendo l'ottava in dodici parti uguali, sulla base della radice dodicesima di 2. Ogni semitono corrisponde a un aumento del 5,9% della frequenza del suono. La scala rappresentata in notazione americana è così composta:

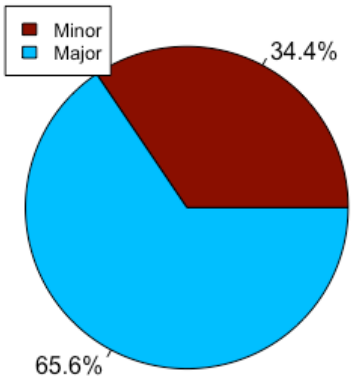
0 = DO,	1 = DO#,	2 = RE,
3 = RE#,	4 = MI,	5 = FA,
6 = FA#,	7 = SOL,	8 = SOL#,
9 = LA,	10 = LA#,	11 = SI



Si nota come la moda sia il SOL con 656 osservazioni.

Mode

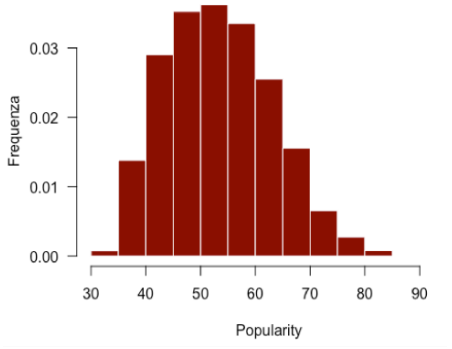
Indica la scala melodica con cui è composto il brano. Il valore 0 indica la scala melodica minore e 1 la maggiore.



Popularity

Variabile che descrive la popolarità della traccia. La popolarità è calcolata da un algoritmo che si basa, nella maggior parte dei casi, sul numero totale di riproduzioni che la traccia ha avuto. Il valore è compreso tra 0 e 100, dove 100 è il più popolare.

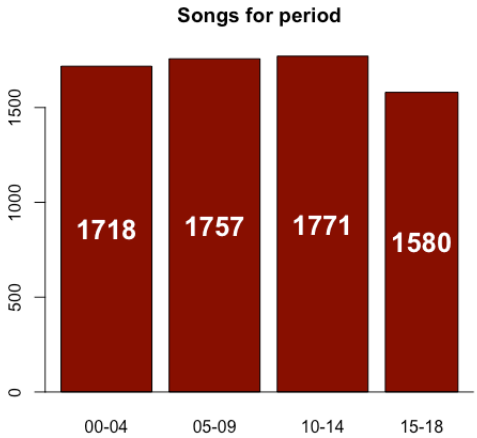
Summary	Popularity
Min.	33
1st Qu.	47
Median	54
Mean	54,04
3rd Qu.	61
Max	91



Year

Indica l'anno di rilascio della canzone.

Si è deciso di creare 4 classi che raccolgano rispettivamente I brani: dal 2000 al 2004, dal 2005 al 2009, dal 2010 al 2014 e dal 2015 al 2019.



Inferenza

Per poter simulare una situazione di varianza nota e valutare intervalli di confidenza e di verifica d'ipotesi assumiamo, solo in questi casi, che la popolazione da cui estraiamo il campione contenga tutte le canzoni tra il 2000 e il 2018.

Intervalli di Confidenza

Per il Teorema del Limite Centrale, avendo una numerosità campionaria molto grande, gli stimatori dei 3 esempi riportati in seguito sotto H_0 si distribuiscono come una $N(0,1)$.

Acousticness —

IC 90% per μ con VAR ignota

Non conoscendo la varianza della popolazione, è stata utilizzata la varianza campionaria corretta. Dopo aver calcolato l'errore campionario si può dire che si è confidenti al 90% che la media della popolazione sia all'interno dell'intervallo:

$$(0{,}249;\ 0{,}259)$$

Danceability —

IC 99% per μ con VAR nota

Conoscendo la media campionaria 0,58863 del campione e la varianza 0,02918 della popolazione si è confidenti al 99% la media della popolazione è all'interno dell'intervallo:

$$(0{,}582;\ 0{,}591)$$

Key —

IC 95% per la PROPORZIONE CAMPIONARIA

Dei 12 semitoni, il SOL che è la chiave maggiormente utilizzata, ha una proporzione campionaria di 0,1141225. Si è quindi confidenti al 95% che la proporzione campionaria di SOL sia compresa all'interno dell'intervallo:

$$(0{,}107;\ 0{,}122)$$

Verifica d'Ipotesi - Media

Per il Teorema del Limite Centrale, avendo una numerosità campionaria molto grande, la statistica test dei 3 casi riportati sotto H_0 si distribuisce come una $N(0,1)$.

Duration —

Verifica d'Ipotesi con α = 0,1 con VAR nota

Si vuole verificare che nella popolazione la durata media delle canzoni non sia superiore a 4min (240000ms).

$$H_0\text{: } \mu = 240000$$

$$H_1\text{: } \mu > 240000$$

$$z_\alpha = \phi^{-1}(1-\alpha) = 1{,}2815$$

$$U = \frac{\bar{X} - \mu_0}{\sigma_0/\sqrt{n}} = -7{,}745$$

$z_\alpha > U$ non si può quindi rifiutare H_0

U	z_α	$p-value$
-7,745	1,2815	1

Mode —

Verifica d'Ipotesi con α = 0.05 con VAR nota

Conosciuta la proporzione campionaria della variabile mode con modalità "minor" p = 0,346616 si vuole verificare che nella popolazione 1/3 delle canzoni abbia scala melodica minore.

H_0: p = 1/3

H_1: p ≠ 1/3

$z_\alpha = \phi^{-1}(\alpha/2) = 1{,}9599$

$U = 2{,}306$

$-z_\alpha < U \nleq z_\alpha$ si rifiuta quindi H_0.

$-z_\alpha$	U	z_α	$p-value$
-1,9599	2,306	1,9599	0,02

Tempo —

Verifica d'Ipotesi con α = 0.01 con VAR ignota

Si vuole verificare che nella popolazione la frequenza dei BPM media sia di 120BPM.

H_0: μ = 120

H_1: $\mu \neq 120$

$z_{\frac{\alpha}{2}} = \phi^{-1}(\frac{\alpha}{2}) = -2{,}5758$

$U = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} = 4{,}3363$

$z_\alpha < U \nleq -z_\alpha$ si rifiuta quindi H_0.

$-z_\alpha$	U	z_α	$p-value$
2,5758	4,3363	-2,5758	0,000015

Verifica d'Ipotesi - Varianza

La statistica test in questo caso si distribuisce come una Chi-quadro con (n-1) gradi di libertà.

Loudness —

Verifica d'Ipotesi con α = 0,05

Si vuole verificare che nella popolazione la varianza del "*volume medio*" delle canzoni sia maggiore a 20.

H_0: $\sigma \geq 20$

H_1: $\sigma < 20$

gdl = (n-1) = 6825

$U = \frac{(n-1)S^2}{\sigma_0^2} = 6704{,}375$

$\chi^2_{1-\alpha;n-1} = 6633{,}97$

$0 < U \nleq \chi^2_{1-\alpha;n-1}$ non si può quindi rifiutare H_0. Il p-value è 0,15 quindi per qualsiasi α < 0.15 non è possibile rifiutare H_0.

U	$\chi^2_{1-\alpha;n-1}$	$p-value$
6704,375	6633,97	0.15

Test di Correlazione

Si è applicato il Test di Correlazione a due coppie di variabili per verificare la presenza di incorrelazione e inferire se la stessa è presente anche nella popolazione. Infatti il test ha come ipotesi:

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 0$$

Acoustic & Loudness

Utilizzando una T di Student con (6826 - 2) g.d.l. si ottiene un valore pari a 0,0008 e quindi incorrelazione.

Il p-value $\simeq 1$ molto alto non ci permette di rifiutare l'ipotesi che non ci sia correlazione nella popolazione.

Loudness & Energy

Utilizzando una T di Student con (6826 - 2) g.d.l. si ottiene un valore di correlazione pari al 75%.

Il p-value $\simeq 0$ per cui si può rifiutare l'ipotesi nulla e affermare che anche nella popolazione le due variabili siano correlate.

Test χ^2 per l'Indipendenza

Si è applicato il Chi Test a una coppia di variabili per verificare la presenza di indipendenza e testare che la stessa sia presente anche nella popolazione. Infatti il test ha come ipotesi:

$$H_0 : \chi^2 = 0$$
$$H_1 : \chi^2 \neq 0$$

Mode & Key

La statistica test è una Chi Quadro con 11 gradi di libertà, pari al numero delle modalità di Mode -1 moltiplicato per il numero delle modalità di Key -1.

Si è ottenuto dal test un p-value = 0 si può affermare quindi che le 2 variabili non sono indipendenti.

Analisi della Varianza

Si vuole verificare che la media di popularity all'interno dei periodi di year del campione sia uguale a quella della popolazione.

	Periods	Residuals
Df	3	6822
SumSq	221585	405943
MeanSq	73862	60
Fvalue	1241	
Pr(>F)	<2e-16	

Il test utilizzato è un test F con 3g.d.l. calcolati come 4 categorie - 1.

Il p-value è la probabilità di osservare un valore di F con 3 g.d.l maggiore del valore di F calcolato. Il valore del p-value prossimo a 0 ci indica che le medie sono significativamente diverse anche nella popolazione perché è stato rifiutato H_0

Modelli

In questo capitolo andremo a costruire un modello univariato e un modello multivariato in cui useremo come target la variabile **popularity**.

Modello Univariato

Nella costruzione del modello univariato utilizzeremo la variabile year come variabile esplicativa.

Si ottiene in questo modo il seguente modello lineare:

$$\textbf{popularity} = 49{,}0890 + 8{,}2041 * \textbf{danceability}$$

Si ricava dal modello che:

- B_0 mostra che nel caso di ballabilità pari a zero, la popolarità sarebbe pari a 49,0890.
- B_1 indica che all'aumentare di un punto del valore di ballabilità, la popolarità aumenta di 8,2041.
- $\mathbb{R}^2$ di questo modello è 0,01767. Per cui questo modello spiega solo l' 1,77% della varianza totale, mentre la restante percentuale è dovuta alla varianza residua.
- Per il ***t-test*** i coefficienti risultano significativamente diversi da 0.

Modello Multivariato

Nell'analisi del modello multivariato si è inizialmente inserito tutte le variabili. È stato riscontrato che solo le variabili: **loudness, speechiness, danceability** e **periods** risultano significative per il ***t-test***, cioè un p-value molto vicino allo 0 che permette di rifiutare l'ipotesi nulla H_0 per cui è stato ricostruito il modello utilizzando solamente queste variabili:

$$\begin{aligned}\textbf{popularity} = {} & 46{,}9275 \\ & + 0{,}21347 * \textbf{loudness} \\ & - 4{,}71597 * \textbf{speechiness} \\ & + 4{,}94123 * \textbf{danceability} \\ & + 2{,}7154 * \textbf{periods05-09} \\ & + 6{,}42601 * \textbf{periods10-14} \\ & + 15{,}07648 * \textbf{periods15-18}\end{aligned}$$

Essendo year una variabile qualitativa divisa in classi, nel modello di regressione viene considerata come in più variabili dummy, una per ciascun livello, nelle quali per ciascuna osservazione verrà assegnato il valore 1 alla variabile del livello di appartenenza e il valore 0 nelle altre.

Si ricava dal modello che:

- La popolarità cresce al crescere di tutte le variabili eccetto che per speechiness
- $\mathbb{R}^2$ risulta 0,3508, per cui il modello spiega il 35,1% della

variabilità di popularity. Il valore di $\mathbb{R}^2$ corretto si attesta attorno a 0,3501. La restante percentuale è dovuta a fattori sconosciuti.

- Per il ***t-test*** i coefficienti risultano significativamente diversi da 0.

- Per il **test F** risulta significativa la relazione tra la variabile dipendente e le variabili esplicative perché il p-value è prossimo a 0.

Si è infine verificato la presenza di eventuale **multicollinearità**. Per fare questo è stato utilizzato il VIF che ci restituisce i seguenti valori:

Variable	VIF
Loudness	1,013317
Speechiness	1,065715
Danceability	1,065617
Periods	1,024969

Avendo per tutte le variabili $VIF < 3$ si può concludere che non è presente un problema di multicollinearità tra le variabili applicative.

Conclusioni

Per quanto l'R^2 ottenuto, sia nel modello univariato che multivariato, non assuma un valore elevato, è stato appurato che l'anno, indicatore della grandiosa diffusione di Spotify il quale ha aumentato la propria adoption, ha influito notevolmente nello spiegare perchè una canzone è popolare.

Per rendere più chiaro questo concetto basti pensare che oggi qualsiasi artista indipendentemente dalla capacità può diventare ascoltatissimo, grazie a questa diffusione che rende popolare chiunque anno dopo anno.

Il modello multivariato ha reso chiaro come le persone apprezzano musica sempre più rumorosa, artificiale, meno acustica e "d'autore", cioè con meno parole e più musicalità.

Il modello ideale per studiare questo fenomeno sarebbe non lineare e andrebbero introdotte altre variabili per provare a spiegare la popolarità della traccia come ad esempio il genere musicale e la casa discografica che ha prodotto il brano.

Bibliografia

- www.kaggle.com
- www. developer.spotify.com