

In [125]:

```
#Michał Wiśniewski, nr 141335
#Informatyka I1
```

In [126]:

```
#Używany zbiór danych pochodzi z Kaggle
#Link - https://www.kaggle.com/leonardopena/top-spotify-songs-from-20102019-by-year

#Jest to baza prawie 600 utworów z bazy danych Spotify, najpopularniejszych w minionej dekadzie z następującymi param
etrami:
#tytuł
#artyста
#rok
#gatunek
#bpm (tempo utworu, ilość uderzeń na minutę)
#Energy (parametr wyliczany za pomocą algorytmu Spotify - im wyższa, tym wyższa energetyczność utworu)
#Danceability (parametr wyliczany za pomocą algorytmu Spotify - im wyższa, tym bardziej nadający się do tańczenia)
#Loudness (db) - poziom głośności
#Liveness (parametr wyliczany za pomocą algorytmu Spotify - im wyższy tym bardziej brzmi jak nagranie na żywo)
#Valence (parametr wyliczany za pomocą algorytmu Spotify - im wyższy tym bardziej pozytywny nastrój)
#duration (długość utworu)

#Acousticness (parametr wyliczany za pomocą algorytmu Spotify - im wyższy tym bardziej akustyczny utwór)
#Speechiness (parametr wyliczany za pomocą algorytmu Spotify - im wyższy tym utwór zawiera sobie więcej mówienia)
#Popularity (popularność w serwisie Spotify)
```

In [127]:

```
library(tidyverse)
library(data.table)
library(readxl)
library(moments)

Sys.setlocale('LC_ALL', 'C')
data <- read.csv("../input/top-spotify-songs-from-2010-2019-by-year/top10s.csv")
```

```
'LC_CTYPE=C;LC_NUMERIC=C;LC_TIME=C;LC_COLLATE=C;LC_MONETARY=C;LC_MESSAGES=C;LC_PAPER=en_US.UTF-
8;LC_NAME=C;LC_ADDRESS=C;LC_TELEPHONE=C;LC_MEASUREMENT=en_US.UTF-8;LC_IDENTIFICATION=C'
```

```
In [128]: #Analiza eksploracyjna -----
data = data[-c(443), ]
# na początek usuwam rekord 443, ponieważ jest błędem - nie zawiera żadnych szczegółowych danych (wszędzie przypisane 0), więc pсуby analizę
```

#1. Najpopularniejsi artyści dekady - z największą liczbą utworów w zbiorze

```
artists <- data[["artist"]]
top_artists <- sort(table(artists), decreasing=TRUE)[1:10]
top_artists
```

	artists	
Katy Perry	Justin Bieber	Maroon 5
17	16	15
Lady Gaga	Bruno Mars	Ed Sheeran
14	13	11
Shawn Mendes	The Chainsmokers	
		11

In [129]:

#2. Najpopularniejsze gatunki na liście (wg. danych Spotify)

```
genres <- data[["top.genre"]]
genres <- sort(table(genres), decreasing=TRUE)[1:5]
genres
```

	genres	
dance	pop	pop
327	60	34
		34
		15
		boy band

In [130]:

```
#3. Tempo utworu - wartości statystyki opisowej  
max(data$bpm)  
min(data$bpm)  
mean(data$bpm)  
median(data$bpm)  
quantile(data$bpm)  
sd(data$bpm)  
skewness(data$bpm)  
kurtosis(data$bpm)
```

#Widac już że rozkład zmiennej tempa utworu jest bardzo skupiony w okolicach średniej, a połowa wartości jest między 100 a 129 bpm. Jest lekko nachylony w lewo oraz ma całkiem wysoką kurtozę.

206

43

118.742524916944

120

**0%: 43 25%: 100 50%: 120 75%: 129 100%: 206**

24.3394974006042

0.735425127996181

4.09243808376602

In [131]:

#Histogram zmiennej tempa utworów, po doborze przedziałów co 10bpm za wyjątkiem skrajnych wartości

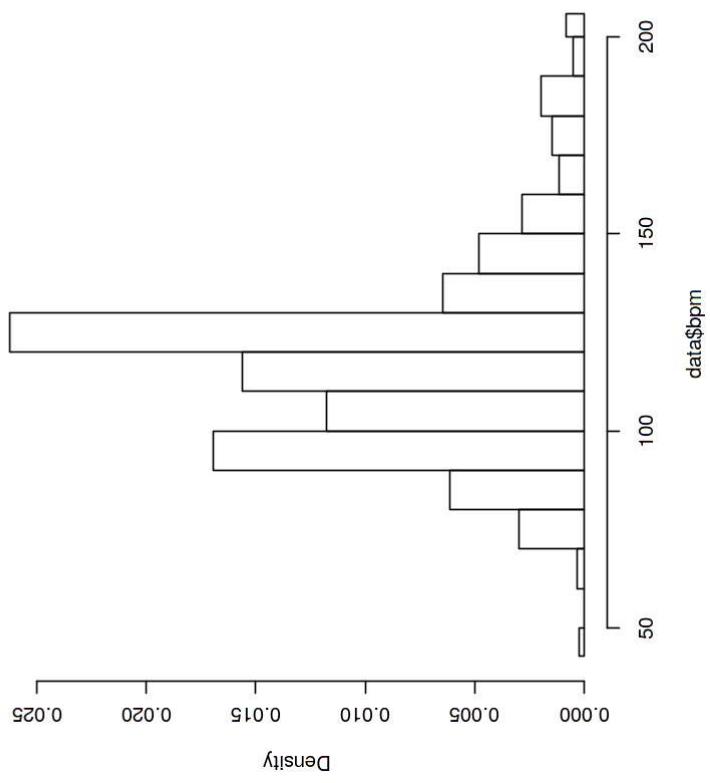
hist(data\$bpm, breaks = c(43, 50, 60, 70, 80, 90, 100, 110, 120, 130, 140, 150, 160, 170, 180, 190, 200, 206),)

#najczesniej wystepujace dokladne wartosci bpm i ile razy wystepuja:

bpms <- data[["bpm"]]  
bpms <- sort(table(bpms), decreasing=TRUE)[1:10]  
bpms

```
bpm  
120 100 128 130 125 126 122 95 124 127  
47 33 29 27 21 20 16 14 12 12
```

Histogram of data\$bpm



In [132]:

```
#4. dynamika (głośność) utworu - wartości statystyki opisowej  
min(data$dB)  
max(data$dB)  
mean(data$dB)  
median(data$dB)  
quantile(data$dB)  
sd(data$dB)  
skewness(data$dB)  
kurtosis(data$dB)
```

#Widac że większość rozkładu jest skupiona w okolicach średniej (połowa wartości między -6 i -4), jest nachylone w prawo i ma wysoką kurtozę. Cechuje się natomiast niskim odchyleniem standartowym, wartości nie różnią się liczbowo o wiele.

-15

-2

-5.48837209302326

-5

0%: -15 25%: -6 50%: -5 75%: -4 100%: -2

1.70465673478451

-0.984065580760153

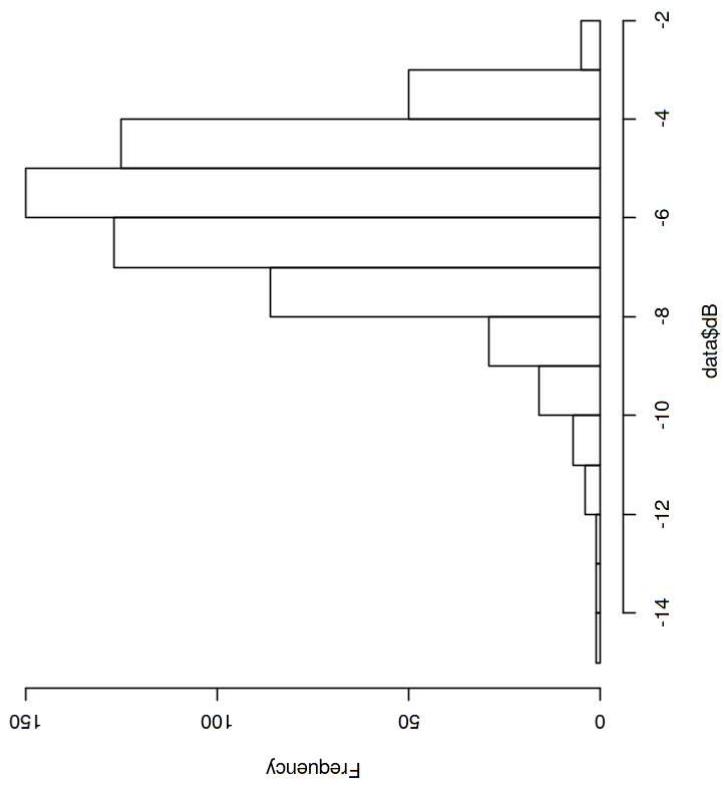
5.3454741242172

In [133]:

```
#szereg rozdzielczy oraz histogram głośności utworów.  
table(data$dB)  
hist(data$dB)
```

```
-15 -13 -12 -11 -10 -9 -8 -7 -6 -5 -4 -3 -2  
1 1 1 4 7 16 29 86 127 150 125 50 5
```

Histogram of data\$dB



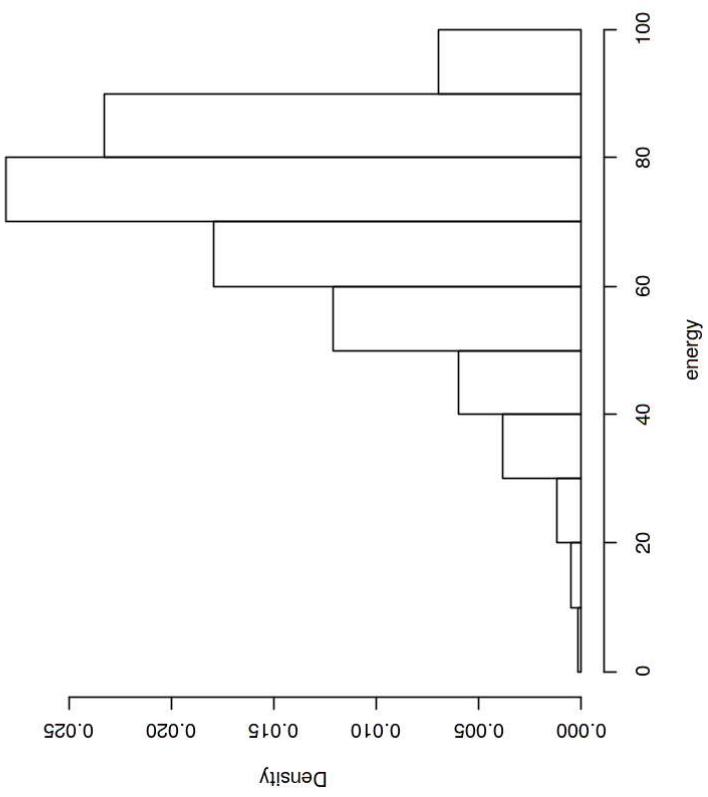
In [134]:

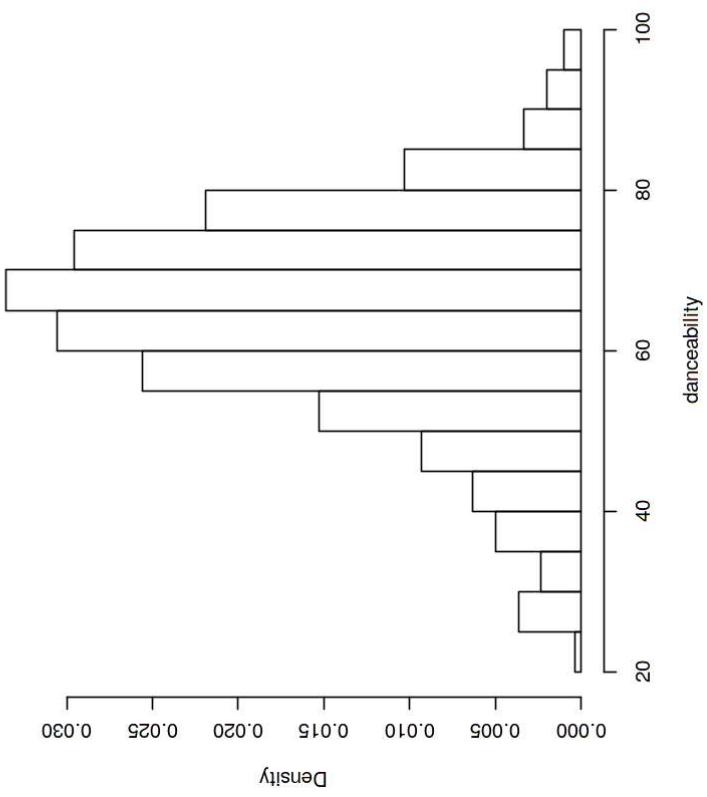
```
#statystyki wykorzystujące algorytm spotify to wartości z zakresu <0,100> dlatego świetnie sprawdza się w nich histogram z domyślnymi przedziałami

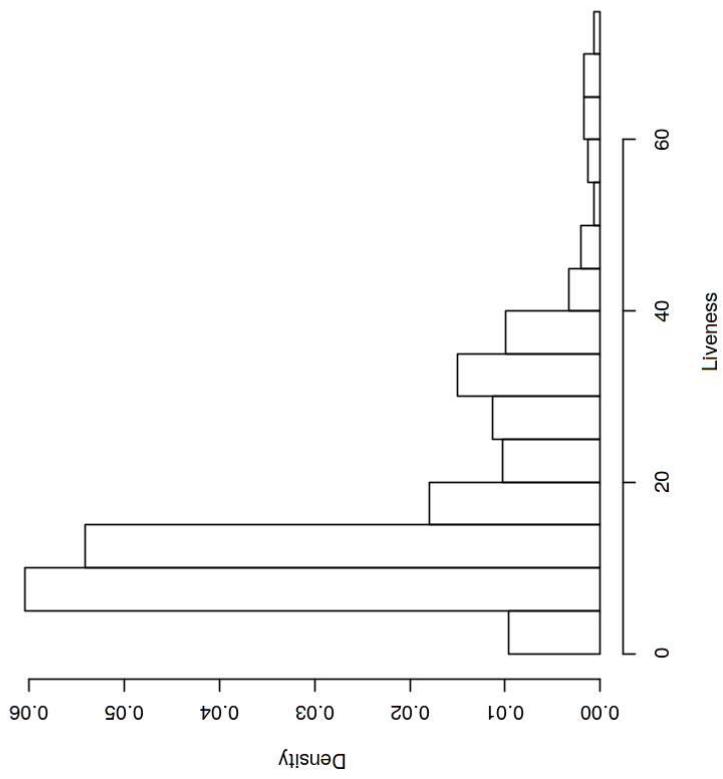
hist(data$nrg, freq = FALSE, xlab = "energy")
hist(data$dnce, freq = FALSE, xlab = "danceability")
hist(data$live, freq = FALSE, xlab = "Liveness")
hist(data$val, freq = FALSE, xlab = "Valence")
hist(data$acous, freq = FALSE, xlab = "Acousticness")
hist(data$spch, freq = FALSE, xlab = "Speechiness")
```

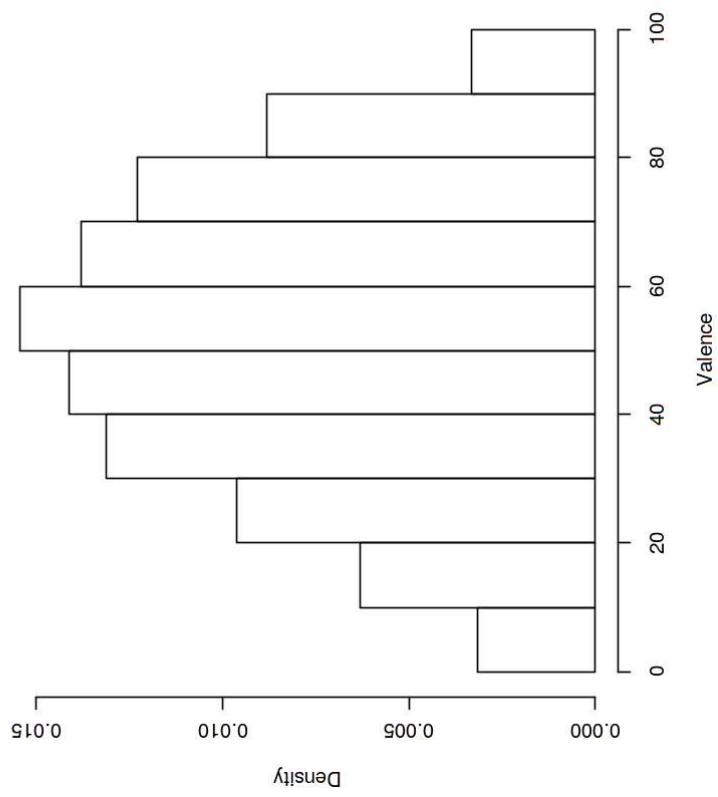
In [134]: #statystyki wykorzystujące algorytm spotify to wartości z zakresu <0,100> dlatego świetnie sprawdza się w nich histogram z domyślnymi przedziałami

```
hist(data$nrg, freq = FALSE, xlab = "energy")
hist(data$dnce, freq = FALSE, xlab = "danceability")
hist(data$live, freq = FALSE, xlab = "Liveness")
hist(data$val, freq = FALSE, xlab = "Valence")
hist(data$acous, freq = FALSE, xlab = "Acousticness")
hist(data$spch, freq = FALSE, xlab = "Speechiness")
```

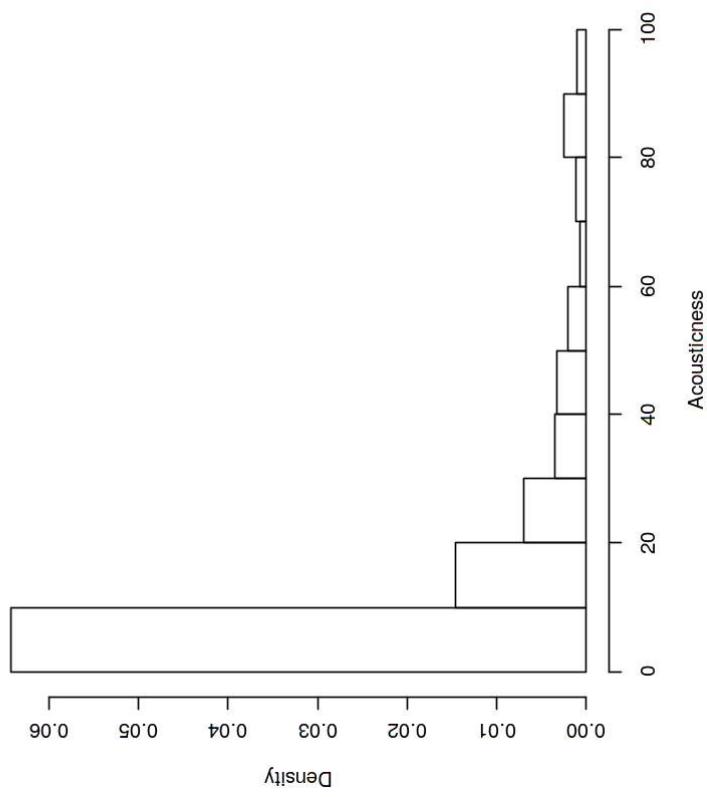
**Histogram of data\$nrg**

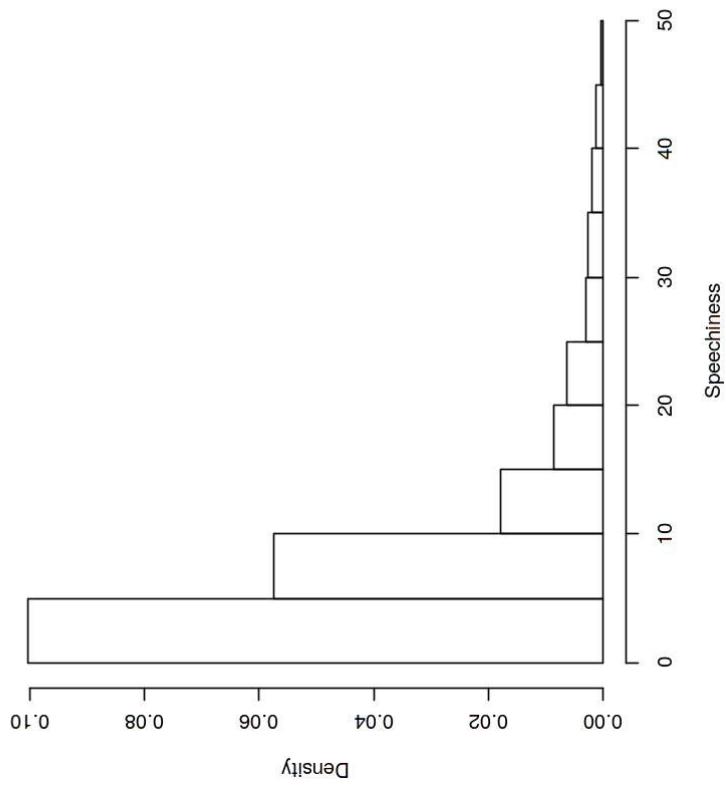
**Histogram of data\$dnce**

**Histogram of data\$live**

**Histogram of data\$val**

### Histogram of data\$acous



**Histogram of data\$spch**

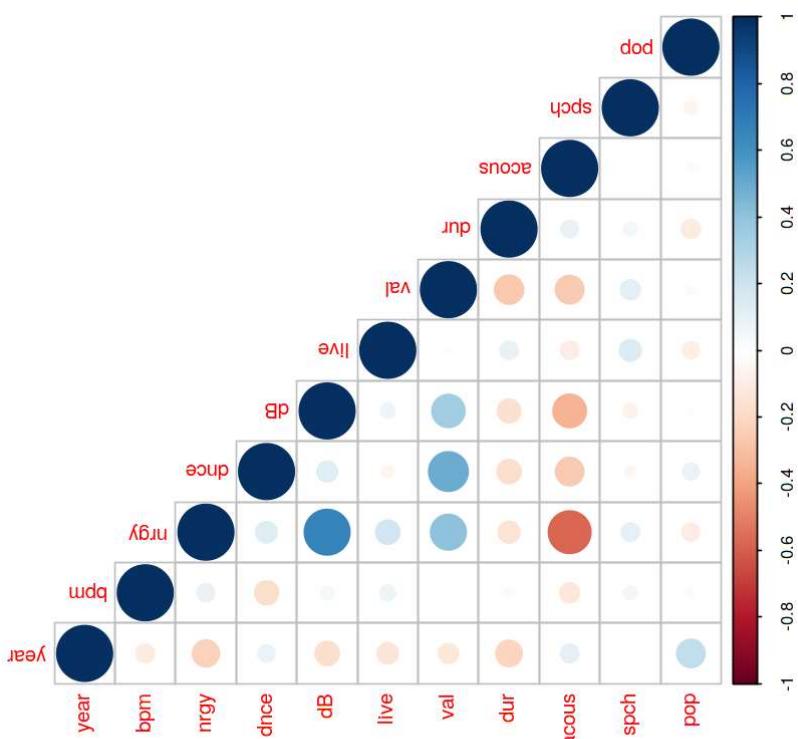
In [135]:

```
#5. Tabela korelacji z liczbowymi parametrami utworów
cor(data[, 5:15])
library(corrplot)
corrplot(cor(data[, 5:15]), type = "lower")
```

A matrix: 11 × 11 of type dbl

	year	bpm	nrgy	dnce	dB	live	val	dur	acous	spc
year	1.0000000	-0.101932410	-0.22529556	0.08527208	-0.17909376	-0.13535237	-0.120510449	-0.21545795	0.102410515	0.C
bpm	-0.10193241	1.000000000	0.09507031	-0.17633102	0.04881749	0.07228434	-0.002491784	-0.02938141	-0.121117530	0.C
nrgy	-0.22529556	0.095070311	1.00000000	0.13738205	0.66362329	0.18007365	0.400945330	-0.14539881	-0.576506537	0.1
dnce	0.08527208	-0.176331018	0.13738205	1.00000000	0.12973696	-0.04051220	0.494927998	-0.17979382	-0.250559655	-0.
dB	-0.17909376	0.048817489	0.66362329	0.12973696	1.00000000	0.06257379	0.3429983408	-0.16841473	-0.349600386	-0.
live	-0.13535237	0.072284342	0.18007365	-0.04051220	0.06257379	1.00000000	0.015080539	0.09864458	-0.099916954	0.1
val	-0.12051045	-0.002491784	0.40094533	0.49492800	0.34298341	0.01508054	1.000000000	-0.26317469	-0.252935836	0.1
dur	-0.21545795	-0.029381407	-0.14539881	-0.17979382	-0.16841473	0.09864458	-0.263174693	1.00000000	0.091917143	0.C
acous	0.10241052	-0.121117530	-0.57650654	-0.25055966	-0.34960039	-0.09991695	-0.252935836	0.09191714	1.00000000	0.C
spch	0.00578947	0.051147757	0.10096518	-0.03775737	-0.06126692	0.14194676	0.118354778	0.05474718	0.001482513	1.C
pop	0.24984077	-0.018128391	-0.09369345	0.08239361	0.01440024	-0.0877029	0.021745572	-0.10570654	0.021837007	-0.





In [136]:

#Najwyraźniejsze korelacje dodatnie - głośność utworu z jego energicznością ( $\theta.66$ ), pozytywny nastrój z tanecznością ( $\theta.49$ ), popularność z rokiem wydania ( $\theta.24$ )

#Najwyraźniejsze korelacje ujemne - akustyczność z energicznością ( $-\theta.57$ ), głośność z głośnością ( $-\theta.34$ ), a także co bardziej interesujące - rok wydania z długością utworu ( $-\theta.21$ ) i energicznością ( $-\theta.22$ )

In [137]:

```
#6. Zmieniające się trendy w muzyce popularnej na przestrzeni dekady - zależności parametrów od roku wydania

library(gplots)
library(ggplot2)

plotmeans(data$bpm ~ data$year, mean.labels = TRUE, digits = 2, n.label = FALSE, barcol = 'white', xlab = "Year",
          ylab = "Mean bpm")

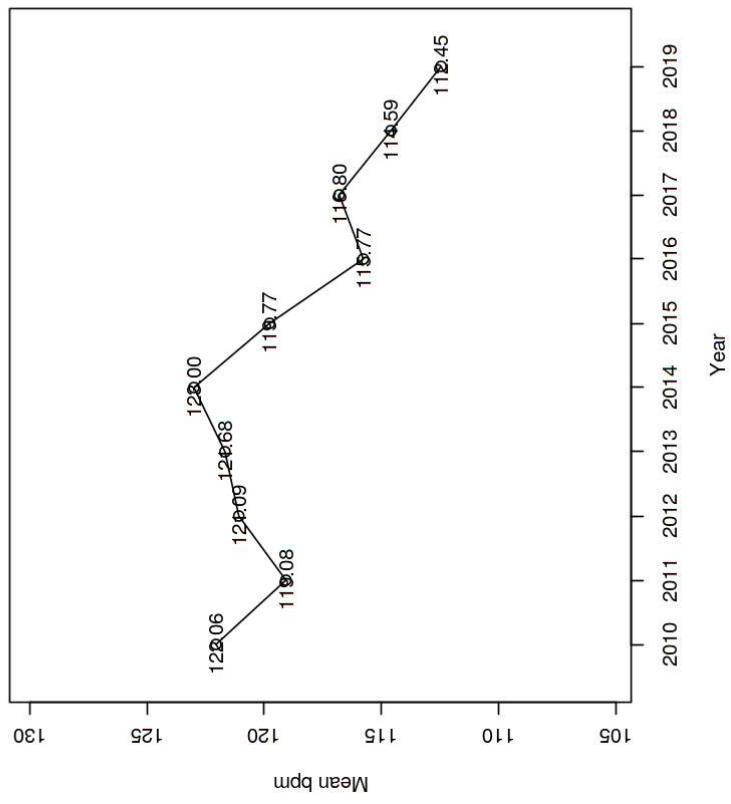
plotmeans(data$nrg ~ data$year, n.label = FALSE, barcol = 'white', xlab = "Year", ylab = "Mean energy")

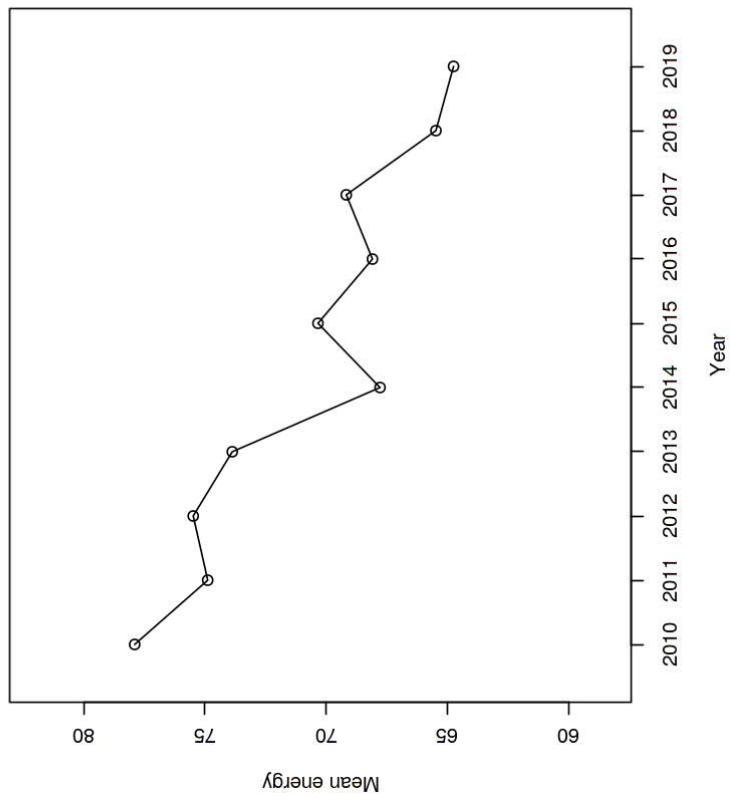
plotmeans(data$dnc ~ data$year, n.label = FALSE, barcol = 'white', xlab = "Year", ylab = "Mean danceability")

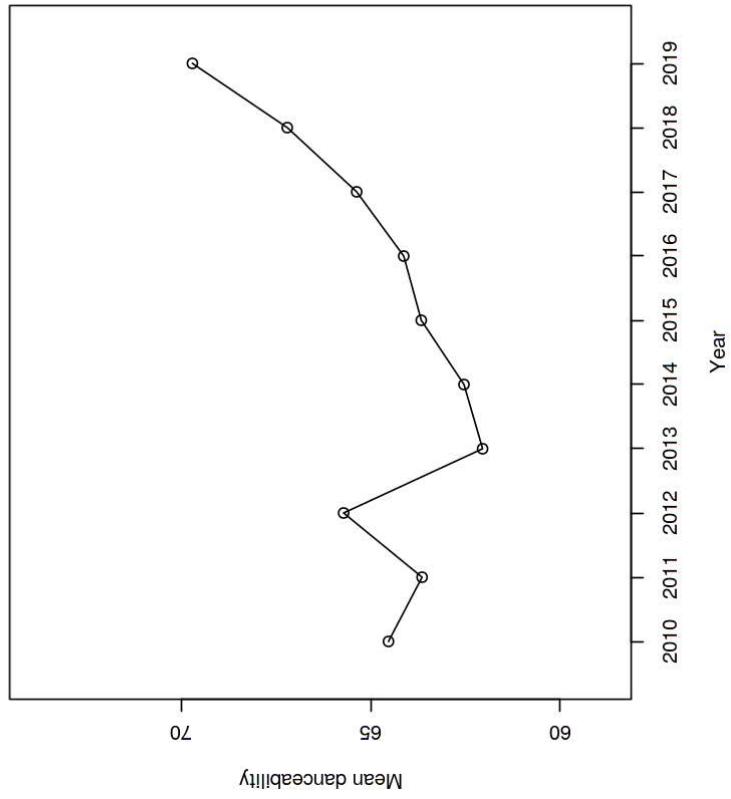
plotmeans(data$dB ~ data$year, n.label = FALSE, barcol = 'white', xlab = "Year", ylab = "Mean Loudness")

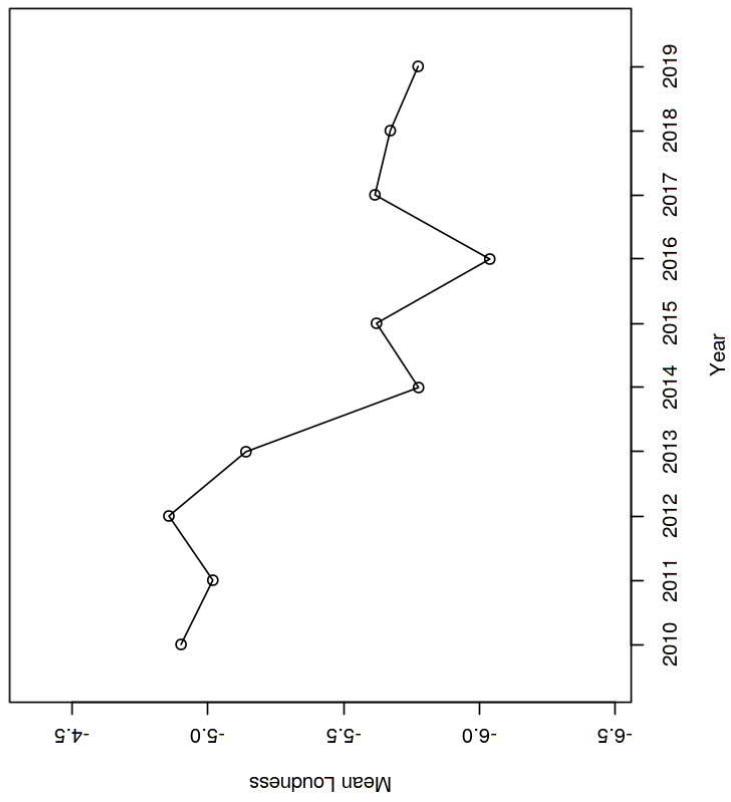
plotmeans(data$dur ~ data$year, n.label = FALSE, barcol = 'white', xlab = "Year", ylab = "Mean duration")

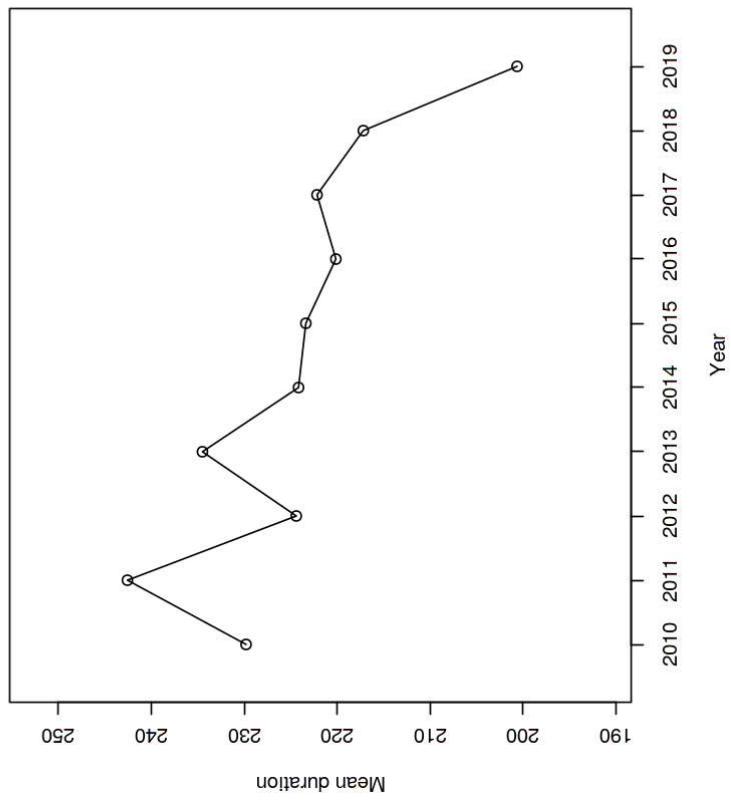
plotmeans(data$val ~ data$year, n.label = FALSE, barcol = 'white', xlab = "Year", ylab = "Mean valence")
```

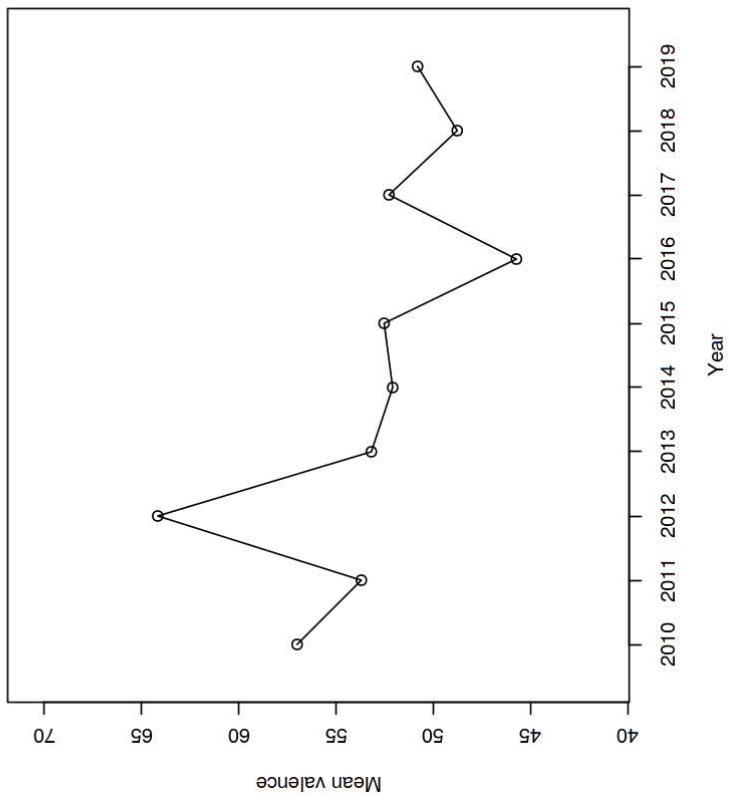












In [138]:

#Wnioski-----

#W ubiegłej dekadzie z czasem wyraźnie malało średnie tempo utworów oraz ich energiczność, przy nieznacznym wzroście taneczności - co ma związek z trendami w muzyce i całej kulturze. Jeszcze ciekawsze są natomiast drastyczne spadki średniej długości utworów oraz głośności - tutaj przyczyny są raczej ekonomiczne związane z rewolucją streamingową. Bardziej opłaca się wydawać krótsze utwory (choćiąż aspekt braku czasu w dzisiejszym świecie może również mieć wpływ), a automatyczne dostosowywanie głośności wprowadzone na większości platform w ostatnich latach sprawia, że nie opłaca się tzw "loudness war" - produkcja muzyki najgöśniej jak to tylko możliwe, comiało miejsce jeszcze kiedyś.

#Spore wahania występuły również w średnim nastroju utworów - zdecydowanie najweselsze hity przypadły na rok 2012, a najsmutniejsze w 2016.

In [139]:

```
#7. Testy statystyczne -----  
# Wybór zbioru danych, który jest biblioteką 28 utworów pewnego użytkownika serwisu. Użytkownik analizuje swój zbior  
i stawia hipotezy.  
  
zbior <- data[c(5, 22, 43, 67, 101, 112, 115, 116, 169, 194, 201, 220, 223, 253, 264, 308, 321, 341, 367, 370, 384  
, 404, 441, 482, 508, 540, 564, 581),]  
zbior  
  
##W związku z tym, że liczba próbki wynosi 28, a odchylenia parametrów całego zbioru danych są znane, przeprowadzone z  
ostaną testy T z poziomem istotności alpha = 0.01
```

A data.frame: 28 × 15

	X	title	artist	top.genre	year	bpm	nrgy	dnce	dB	live	val	dur	acous	spch	pop
	<int>	<fct>	<fct>	<fct>	<int>										
5	5	Just the Way You Are	Bruno Mars	pop	2010	109	84	64	-5	9	43	221	2	4	78
22	22	Whataya Want from Me	Adam Lambert	australian pop	2010	186	68	44	-5	6	45	227	1	5	66
43	43	Hard	Rihanna	barbadian pop	2010	182	75	31	-4	65	16	251	1	11	57
67	67	Party Rock Anthem	LMFAO	dance pop	2011	130	74	75	-4	27	35	262	2	16	72
101	101	Moment 4 Life - Album Version (Edited)	Nicki Minaj	dance pop	2011	130	88	50	-4	22	37	279	39	38	28
112	112	Love You Like A Love Song	Selena Gomez & The Scene	dance pop	2012	117	68	86	-4	7	92	188	8	5	76
115	115	Stronger (What Doesn't Kill You)	Kelly Clarkson	dance pop	2012	116	94	56	-4	11	68	222	5	5	74
116	116	Try	P!nk	dance pop	2012	104	63	67	-7	9	55	248	0	3	74
169	169	Let Me Love You (Until You Learn To Love Yourself)	Ne-Yo	dance pop	2013	125	68	66	-7	37	25	252	25	4	70
194	194	What About Love	Austin Mahone	dance pop	2013	100	78	63	-7	4	28	203	0	4	54
201	201	How Ya Doin? (feat. Missy Elliott)	Little Mix	dance pop	2013	201	95	36	-3	37	51	211	9	48	50
220	220	Treasure	Bruno Mars	pop	2014	116	69	87	-5	32	94	179	4	4	77
223	223	Pompeii	Bastille	metropolis	2014	127	72	68	-6	27	57	214	8	4	73
253	253	Chandelier	Sia	australian dance	2014	174	78	29	-3	7	62	216	2	7	56
264	264	Sheezus	Lily Allen	dance pop	2014	130	50	78	-7	11	39	235	15	4	37
308	308	Heartbeat Song	Kelly Clarkson	dance pop	2015	149	80	49	-4	6	48	199	1	5	69
321	321	I Really Like You	Carly Rae Jepsen	canadian pop	2015	122	81	62	-5	22	60	205	1	4	66

	X	title	artist	top.genre	year	bpm	nrgy	dnce	dB	live	val	dur	acous	spch	pop
	<int>	<fct>	<fct>	<fct>	<int>										
341	341	Heroes (we could be)	Alesso	big room	2015	126	75	52	-4	24	35	210	4	6	56
367	367	Don't Let Me Down	The Chainsmokers	electropop	2016	160	87	53	-5	14	42	208	16	17	81
370	370	This Is What You Came For	Calvin Harris	dance pop	2016	124	93	63	-3	15	47	222	20	3	80
384	384	Shout Out to My Ex	Little Mix	dance pop	2016	126	75	77	-4	11	80	246	2	9	77
404	404	Runnin' (Lose It All)	Naughty Boy	tropical house	2016	139	85	32	-6	48	8	213	1	8	69
441	441	Picky - Remix	Joey Montana	latin	2016	186	81	70	-3	37	69	225	9	7	29
483	483	Strip That Down (feat. Quavo)	Liam Payne	dance pop	2017	106	50	87	-5	8	55	202	20	5	69
509	509	One Kiss (with Dua Lipa)	Calvin Harris	dance pop	2018	124	86	79	-3	8	59	215	4	11	86
541	541	Anywhere	Rita Ora	dance pop	2018	107	80	63	-4	10	32	215	4	6	74
565	565	Filthy	Justin Timberlake	dance pop	2018	97	58	75	-6	25	65	294	4	14	62
582	582	Good as Hell (feat. Ariana Grande) - Remix	Lizzo	escape room	2019	96	89	67	-3	74	48	159	30	6	90

In [140]:

```
# Test statystyczny 1. Użytkownikowi wydaje się że średnie tempo wszystkich utworów z danych może być wyższe, ponieważ  
ż średnia utworów które słucha wyszła znacznie wyższa  
#H0 średnie tempo = 118, 75 (hipoteza prawdziwa)  
#h1 średnie tempo > 118,75  
alpha <- 0.01  
nrow(zbior)  
n <- nrow(zbior)  
stdev <- sd(data$bpm)  
T <- (mean(zbior$bpm) - mean(data$bpm)) / stdev * sqrt(n)  
  
qt(alpha, n - 1)
```

```
abs(T) > abs(qt(alpha, n-1))
```

#Wartość statystyki jest większa od wartości krytycznej, czyli hipoteza  $H_0$  zostaje odrzucona. Popełniono zatem błąd I rodzaju, okazało się że żbior utworów użytkownika był niereprezentatywny dla ogólnej bazy - w jego guscie są szyszce kawałki.

28

2.983164855222676

-2.47265991195601

TRUE

In [141]:

```
#Test statystyczny 2. Użytkownikowi wydaje się, że średnia długość utworów może być krótsza niż podana rzeczywista.  
#H0 średnia długość = 224.67 sekundy (hipoteza prawdziwa)  
#H1 Średnia długość < 224.67 sekundy
```

```
alpha <- 0.01  
n <- nrow(zbior)  
stdev <- sd(data$dur)  
T <- (mean(zbior$dur) - mean(data$dur)) / stdev * sqrt(n)  
  
qt(alpha, n - 1)  
abs(T) > abs(qt(alpha, n-1))
```

#wartość statystyki testowej nie przekracza wartości krytycznej. Hipoteza  $H_0$  zasadnie nie zostaje odrzucona.

-0.386119867105115

-2.47265991195601

FALSE

In [142]:

#Test statystyczny 3. Tym razem użytkownik usłyszał na forum, że średnia głośność najpopularniejszych utworów w ostatniej dekadzie spadła aż do -6 dB, postanowił zakwestionować to na podstawie swojej biblioteki, która wydawała mu się głosniejsza

#H0 średni poziom dB = -6 (hipoteza fałszywa)

#H1 średni poziom dB > -6

```
alpha <- 0.01
n <- nrow(zbior)
stdev <- sd(data$dB)
T <- (mean(zbior$dB) - (-6)) / stdev * sqrt(n)

T

qt(alpha, n - 1)
abs(T) > abs(qt(alpha, n-1))
mean(zbior$dB)
mean(data$dB)
```

#Wartość statystyki jest większa od wartości krytycznej, czyli hipoteza  $H_0$  zostaje odrzucona. Słusznie, ponieważ średnia głośność jest trochę większa niż usłyszał na forum (-5,48, a nie -6)

4.21276896435288

-2.47265991195601

TRUE

-4.64285714285714

-5.48837209302326