Sveučilište u Zagrebu
Prirodoslovno-matematički fakultet
Matematički odsjek

Ivka Ćaćić, Laura Horvat, Goran Ivanković, Ivan Ljubetić, Maja Pavičić

Najslušanije pjesme na Spotifyju u razdoblju 2010. – 2019.

Zagreb, lipanj 2022.

Sadržaj

1 Sažetak

U ovome radu promatrat ćemo popis 100 najslušanijih pjesama na Spotifyju svake godine između 2010. i 2019. Osim naziva pjesmi i autora, na popisu su navedeni još mnogi zanimljivi podaci poput glasnoće, žanra, tempa i drugih. Mi ćemo proučavati razdiobe tih varijabli, promatrati kako se vrijednosti mijenjaju kroz godine te pokušati pronaći zavisnosti.
Na temelju opisne statistike zaključit ćemo koje su nam hipoteze zanimljive za testiranje, pa ćemo odgovarajuće testove provesti u odjeljku Inferencijalna statistika. Na kraju ćemo obuhvatiti sve rezultate testova i donijeti svoj zaključak o najslušanijim pjesmama.

2 Uvod

Podaci su preuzeti sa stranice *Kaggle*, a obrađeni u R-u i Pythonu. Glavni je predmet našeg istraživanja skup podataka koji opisuje *sto najslušanijih pjesama na Spotifyju svake godine između 2010. i 2019.* Međutim, kao referentni set podataka za predviđanje očekivanih vrijednosti koristit ćemo tablicu *sto najslušanijih pjesama na Spotifyju*.
Izgledi setova podataka, odnosno nazivi stupaca s pripadnim opisom prikazani su u Tablici 2.1. Svojstva navedena u tablici računaju se pomoću ustaljenih Spotifyjevih algoritama, u čije detalje nećemo ulaziti.

STUPAC	OPIS
title	naslov pjesme
artist	ime izvođača
genre	žanr pjesme
year released	godina kad je pjesma izašla
added	datum kad je pjesma dodana na Spotifyjevu top listu
bpm	tempo pjesme (broj taktova u minuti)
nrgy	mjera intenziteta i energičnosti pjesme, cijeli broj između 0 i 100
dnce	koliko je pjesma plesna, mjereno na temelju tempa te čujnosti i stabilnosti ritma, cijeli broj između 0 i 100
dB	glasnoća pjesme u decibelima
live	vjerojatnost·100% da je snimka uživo, cijeli broj između 0 i 100
val	mjera pozitivnosti pjesme, cijeli broj između 0 i 100. Pjesme s visokim vrijednostima generalno izazivaju pozitivne emocije
dur	trajanje pjesme u sekundama
acous	mjera akustičnosti pjesme, cijeli broj između 0 i 100
spch	mjera sličnosti pjesme i recitacije, cijeli broj između 0 i 100
pop	popularnost pjesme temeljena na broju slušanja u relevantnoj godini
top year	godina kad je bila na top listi, cijeli broj između 2010 i 2019
artist type	tip izvođača - Solo, Duo, Trio, Band/Group

Tablica 2.1: Imena i opisi stupaca u skupovima podataka

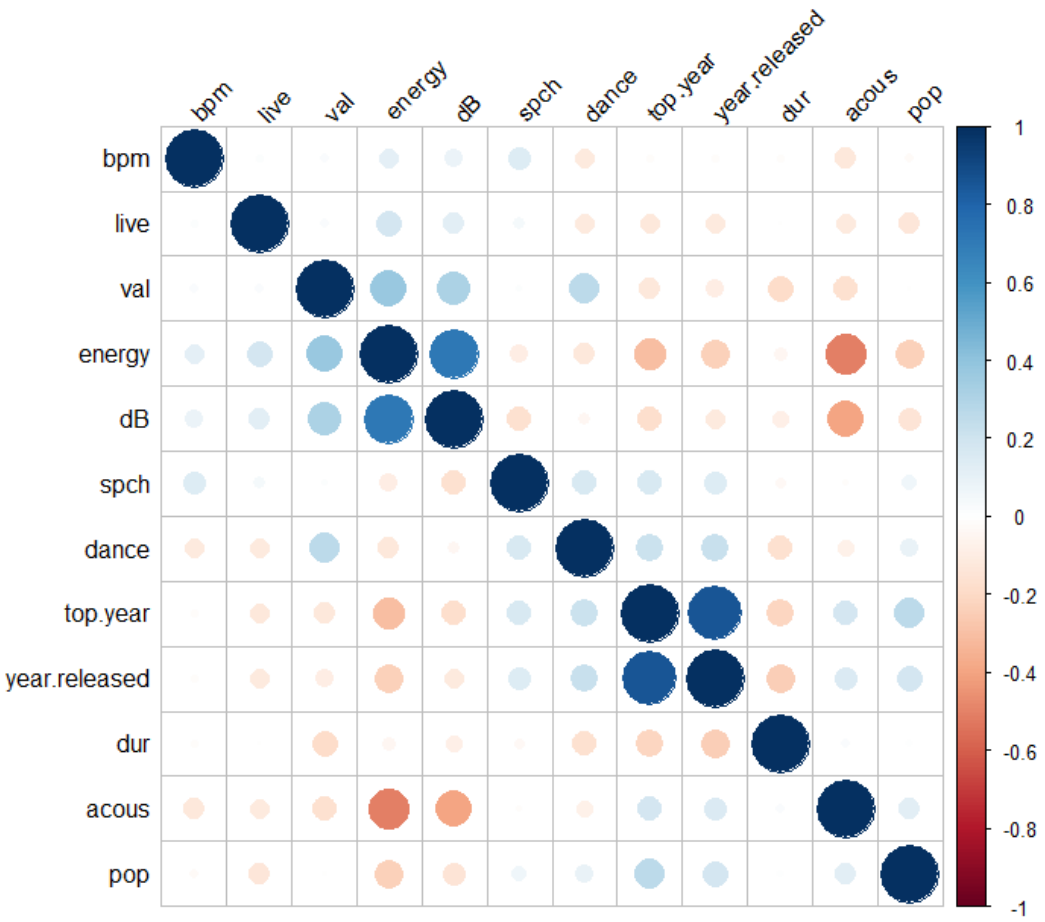
Na Slici 2.1 prikazano je prvih nekoliko pjesama i njihovih svojstava. Podaci u tablici sortirani su uzlazno najprije prema promatranoj godini, a unutar svake godine po izvođaču. Iz tog razloga nemamo podatke o poretku na samoj rang-listi, no svakako i bez toga ne nedostaje podataka za proučavanje.

title	artist	top genre	year released	added	bpm	nrgy	dnce	dB	live	val	dur	acous	spch	pop	top year	artist type
STARSTRUKK (feat. Katy Perry)	3OHI3	dance pop	2009.0	2022-02-17	140.0	81.0	61.0	-6.0	23.0	23.0	203.0	0.0	6.0	70.0	2010.0	Duo
My First Kiss (feat. Ke$ha)	3OHI3	dance pop	2010.0	2022-02-17	138.0	89.0	68.0	-4.0	36.0	83.0	192.0	1.0	8.0	68.0	2010.0	Duo
I Need A Dollar	Aloe Blacc	pop soul	2010.0	2022-02-17	95.0	48.0	84.0	-7.0	9.0	96.0	243.0	20.0	3.0	72.0	2010.0	Solo
Airplanes (feat. Hayley Williams of Paramore)	B.o.B	atl hip hop	2010.0	2022-02-17	93.0	87.0	66.0	-4.0	4.0	38.0	180.0	11.0	12.0	80.0	2010.0	Solo
Nothin' on You (feat. Bruno Mars)	B.o.B	atl hip hop	2010.0	2022-02-17	104.0	85.0	69.0	-6.0	9.0	74.0	268.0	39.0	5.0	79.0	2010.0	Solo

Slika 2.1: Izgled seta podataka

3 Opisna statistika

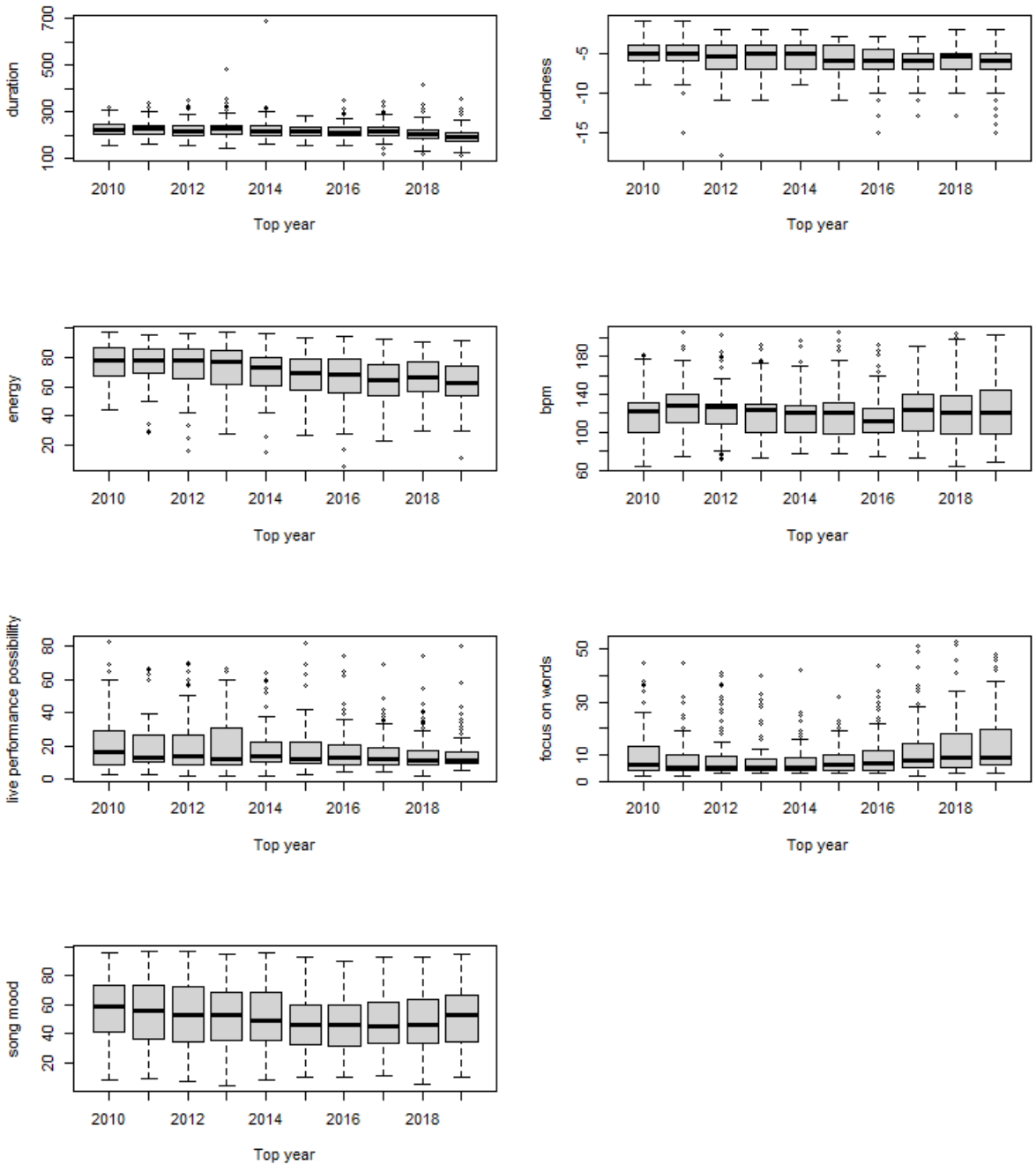
3.1 Korelacija varijabli



Slika 3.1: Dijagram korelacije numeričkih varijabli

Iz ovog dijagrama vidimo da su značajno pozitivno korelirani energija i glasnoća te godina objave pjesme i godina najveće popularnosti. Korelaciju energije i glasnoće kasnije ćemo dodatno promatrati. S druge strane, iz popisa pjesama možemo vidjeti da se razlika godine objave i najveće popularnosti najčešće razlikuju za jednu ili nula godina. Ti podaci zaista imaju smisla jer očekujemo da će pjesma najviše puta biti slušana unutar godine dana od izlaska. Dodatne testove o toj razlici godina stoga nema smisla provoditi.

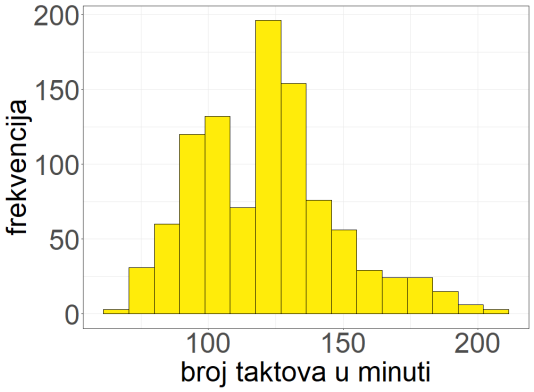
3.2 Grafovi po godinama



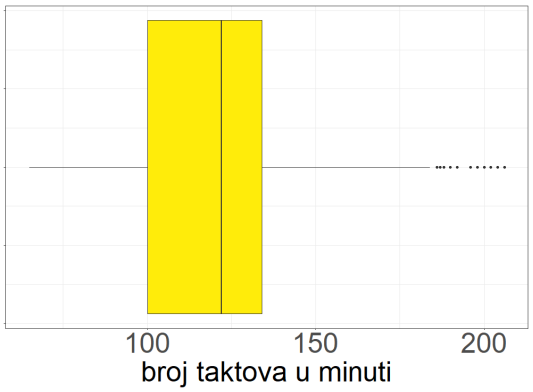
Slika 3.2: Pravokutni dijagrami podataka raspodijeljeni po godinama

Iz pravokutnih dijagrama po godinama naslućujemo da su varijance po godinama slične za sva obilježja, što, ukoliko se distribucije promatranih obilježja pokažu normalne, možemo provjeriti F-testom te zatim testirati jesu li očekivane vrijednosti jednake za različite godine.

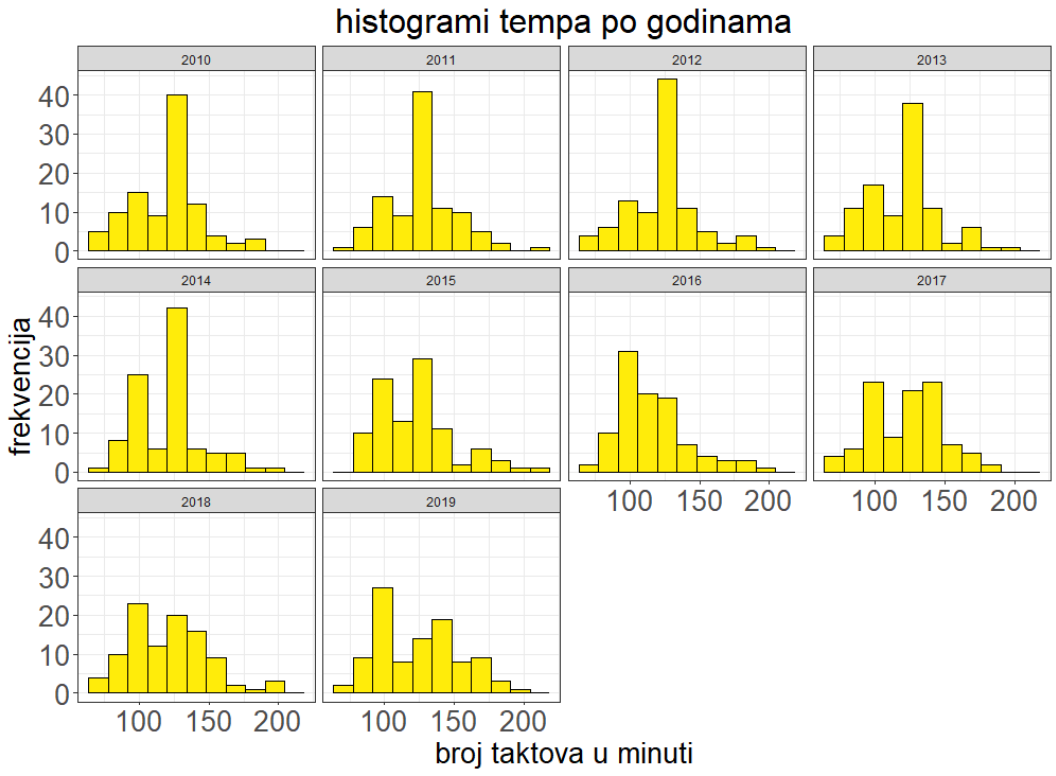
3.3 Tempo



Slika 3.3: Histogram tempa



Slika 3.4: Boxplot tempa



Slika 3.5: Histogrami tempa po godinama

min	q_L	m	q_U	max	srednja vrijednost
65	100	122	134	206	121.26

Tablica 3.1: Karakteristična petorka za tempo

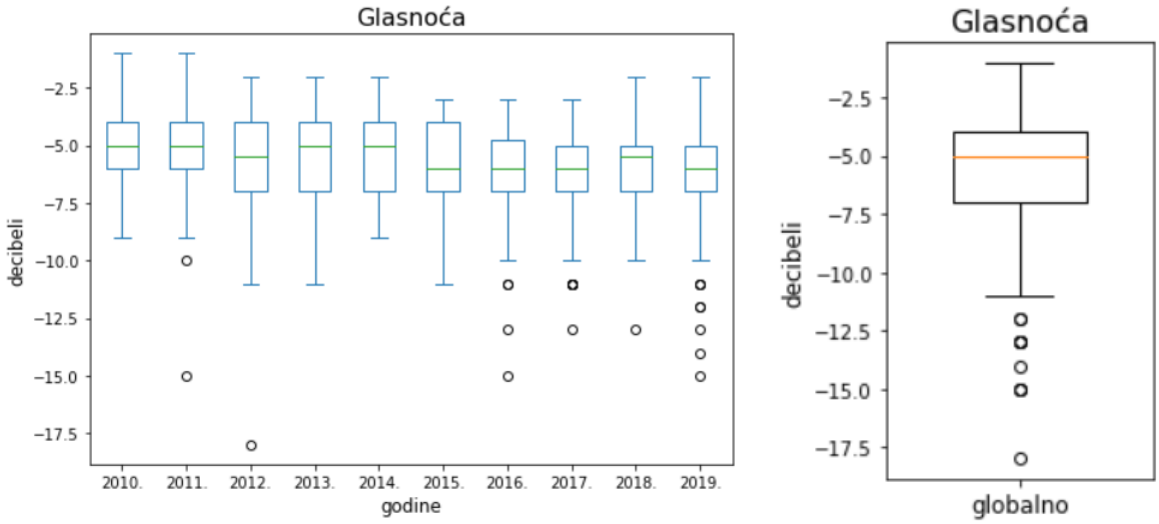
S obzirom na veliku količinu podataka ($n = 1000$), histogram tempa je vrlo nepravilan pa već sada pretpostavljamo da podaci za tempo neće pripadati nekoj od poznatijih distribucija. Pretpostavljamo da bi se moglo raditi o bimodalnoj distribuciji. Lillieforsovim ćemo testom pokušati isključiti mogućnost pripadnosti ove slučajne varijable normalnoj distribuciji.
Distribucija tempa unutar pojedinih godina također ne izgleda normalno. Histogrami se međusobno dosta razlikuju, vidimo da je ranijih godina raspon tempa $122 - 135$ bio dominantan, dok je zadnjih godina njegova popularnost pala.

3.4 Glasnoća

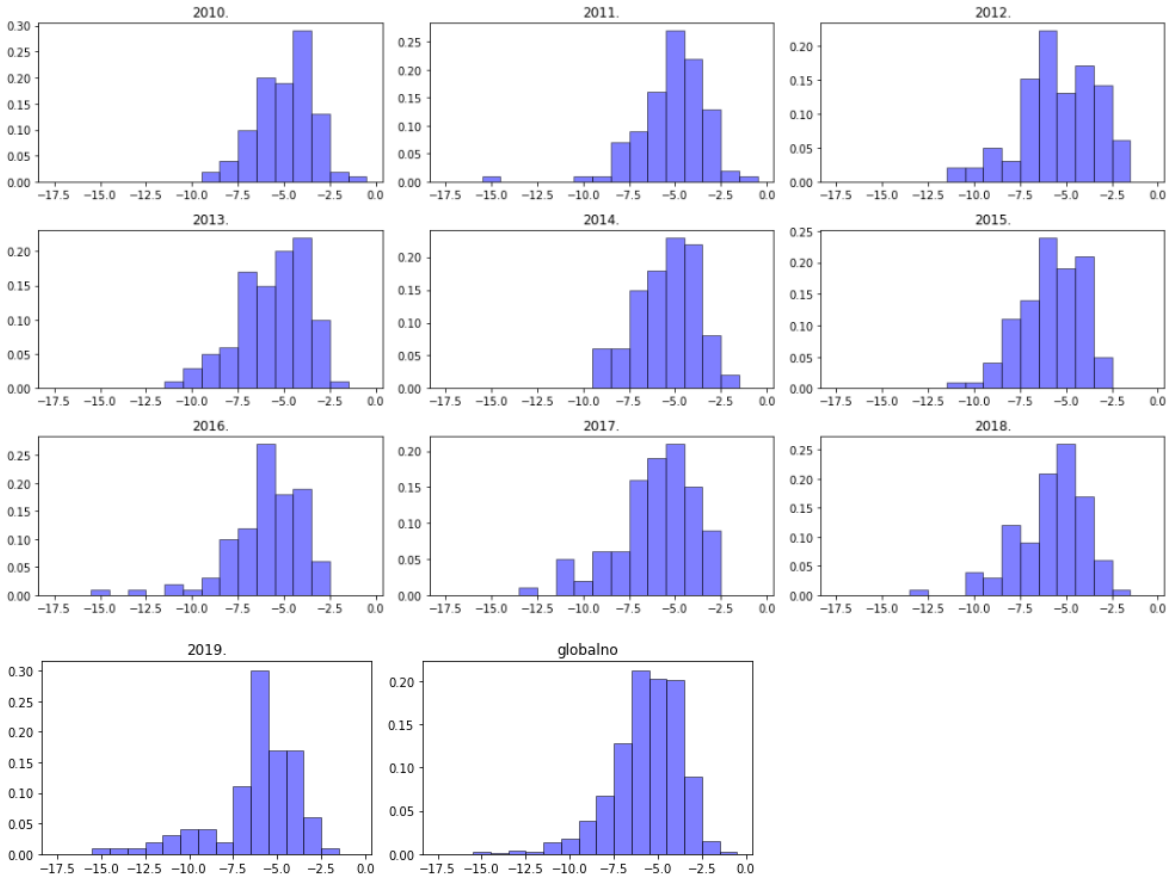
U nastavku prilažemo tablicu karakterističnih petorki za podatke o glasnoći, izračunate za svaku godinu posebno te naposljetku za sve godine zajedno.

Godina	Minimum	Donji kvartil	Medijan	Gornji kvartil	Maksimum
2010	-9.0	-6.0	-5.0	-4.00	-1.0
2011	-15.0	-6.0	-5.0	-4.00	-1.0
2012	-18.0	-7.0	-5.5	-4.00	-2.0
2013	-11.0	-7.0	-5.0	-4.00	-2.0
2014	-9.0	-7.0	-5.0	-4.00	-2.0
2015	-11.0	-7.0	-6.0	-4.00	-3.0
2016	-15.0	-7.0	-6.0	-4.75	-3.0
2017	-13.0	-7.0	-6.0	-5.00	-3.0
2018	-13.0	-7.0	-5.5	-5.00	-2.0
2019	-15.0	-7.0	-6.0	-5.00	-2.0
globalno	-18.0	-7.0	-5.0	-4.00	-1.0

Slika 3.6: Tablica karakterističnih petorki glasnoće



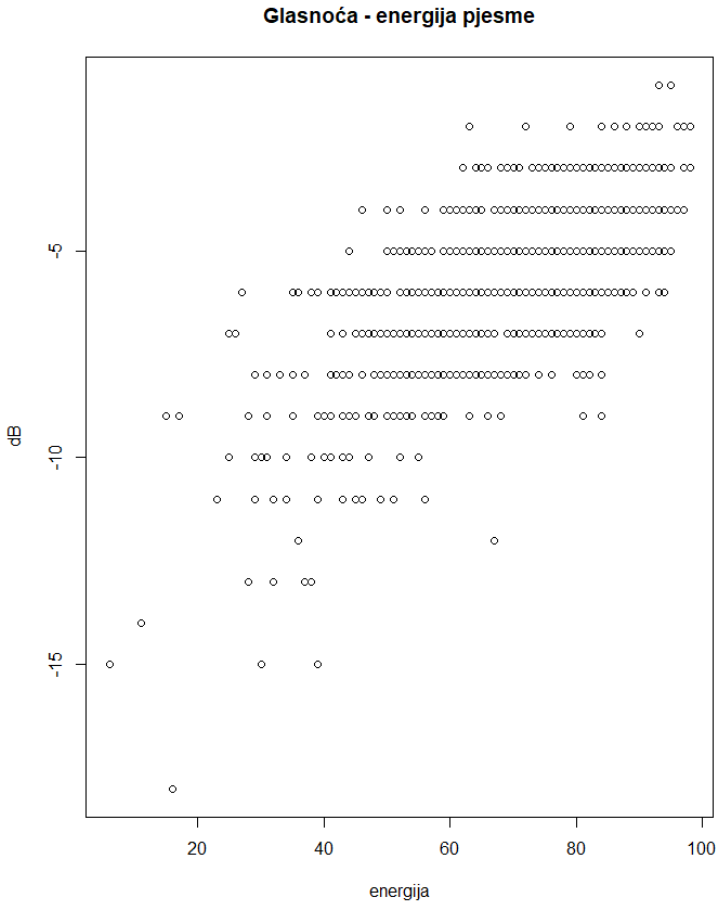
Slika 3.7: Boxplot dijagrami glasnoće po godinama i globalno



Slika 3.8: Histogrami glasnoće kroz godine i globalno

Iz boxplot dijagrama sa Slike 3.7 i histograma na Slici 3.8 možemo primijetiti da je glasnoća većine pjesama svake godine, ali i globalno, između -6.5 i -4.5 dB. Kasnije ćemo testirati je li normalno distribuirana, što se čini opravdanim sudeći po obliku histograma za ukupne podatke.

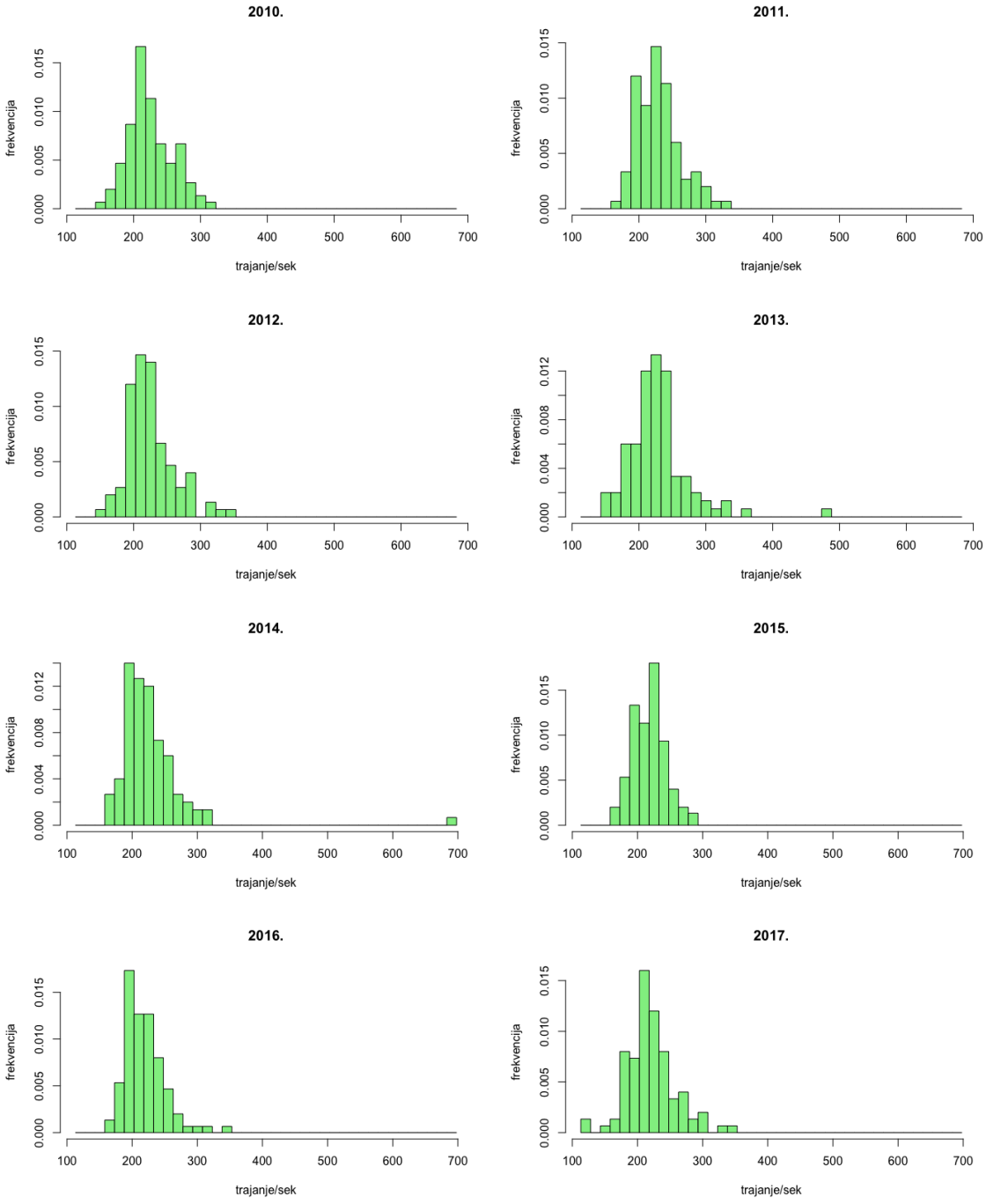
Promotrimo sada odnos između glasnoće i energije pjesama.

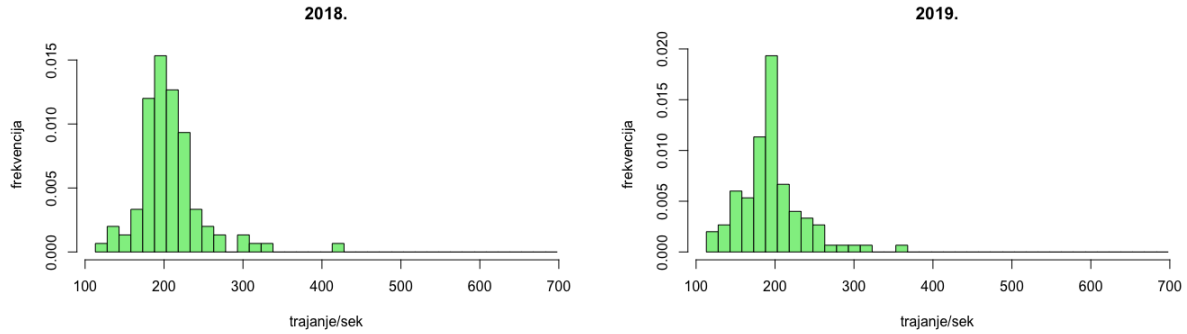


Slika 3.9: Graf energije i glasnoće pjesama

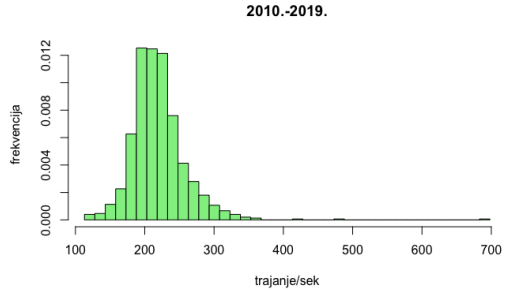
Na grafu se vidi da su pjesme s većom energijom uglavnom glasnije. Možemo naslutiti linearnu zavisnost podataka koju ćemo kasnije provjeriti linearnom regresijom.
Moguće je da je tu ovisnost uzrokovao i sam algoritam. Naime, možda je Spotify algoritam za određivanje energije uvelike u obzir uzimao glasnoću pjesme. Ipak, dublje u ovu temu nećemo zalaziti.

3.5 Trajanje

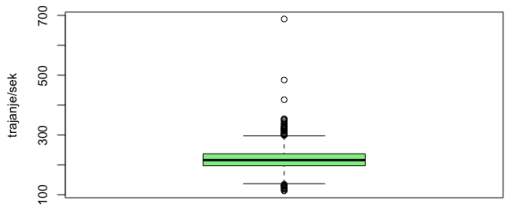

2010.
2011.
2012.
2013.
2014.
2015.
2016.
2017.
frekvencija
trajanje/sek



Slika 3.10: Histogrami trajanja pjesama po godinama



Slika 3.11: Histogram trajanja



Slika 3.12: Boxplot trajanja

min	q_L	m	q_U	max	srednja vrijednost
113	197	216	237	688	220.406

Tablica 3.2: Karakteristična petorka za trajanje

Iz boxplota trajanja pjesama po godinama vidimo da su varijance slične, a histogrami izgledaju približno normalno distribuirani. Ove ćemo teze stoga testirati F-testovima i Lillieforsovim testovima pripadnosti nekoj normalnoj distribuciji pa u slučaju da ne odbacimo nulte hipoteze, možemo provesti i ANOVA-u kako bismo testirali razlikuju li se statistički značajno očekivana trajanja pjesama po godinama.
Najmanji i najveći outlieri podataka su pjesme "Not a Bad Thing" iz 2014. Justina Timberlakea koja traje 688 sekundi te "Old Town Road" Lil Nas Xa iz 2019. u trajanju 113 sekundi. Za znatiželjnije čitatelje navodimo cijeli popis.

	title	dur
1	Gucci Gang	124
2	Jocelyn Flores	119
3	Mine	131
4	Moonlight	135
5	changes	122
6	Thotiana	129
7	emotions	131
8	Old Town Road	113
9	Panini	115
10	Ransom	131
11	My Type	126

	title	dur
1	The Time (Dirty Bit)	308
2	Riverside	321
3	Lighters	304
4	Holocene	337
5	All Of The Lights	300
6	The Edge Of Glory	321
7	Mercy	329
8	Swimming Pools (Drank) - Extended Version	314
9	m.A.A.d city	350
10	Anna Sun	321
11	Lose Yourself to Dance (feat. Pharrell Williams)	354
12	Holy Grail	338
13	Mirrors	484
14	Suit & Tie (feat. Jay-Z)	326
15	All I Want	306
16	Same Love (feat. Mary Lambert)	319
17	Drunk in Love (feat. Jay-Z)	323
18	Not a Bad Thing	688
19	Stolen Dance	314
20	Collard Greens	300
21	Low Life (feat. The Weeknd)	314
22	Somebody Else	348
23	Redbone	327
24	Passionfruit	299
25	Little Dark Age	300
26	Bad and Boujee (feat. Lil Uzi Vert)	343
27	Slippery (feat. Gucci Mane)	304
28	The Greatest Show	302
29	MotorSport	303
30	Te Boté - Remix	418
31	Powerglide (feat. Juicy J) - From SR3MM	332
32	SICKO MODE	313
33	China	302
34	Bohemian Rhapsody - Remastered 2011	354
35	SICKO MODE	313

Slika 3.13: Popis outliera

3.6 Žanrovi po godinama

Slika 3.14: Dijagrami zastupljenosti top 5 žanrova po godinama

Na grafu su prikazani udjeli pjesama određenog žanra među svim pjesama te godine. Jasno je vidljivo da žanr dance pop dominira svake godiname, no možemo naslutiti da se trendovi mijenjaju. Naime, na drugom je mjestu po pojavljivanju najčešće pop (čak 4 puta), ali su se pojavili i alt hip hop i art pop. U svakom slučaju, vidimo da se top 5 žanrova uvelike mijenjaju iz godine u godinu, a kasnije ćemo slično pokazati i χ^2-testom homogenosti.

4 Inferencijalna statistika

4.1 Z-testovi usporedbe očekivanja za tempo, trajanje i glasnoću

U ovom smo dijelu proveli tri Z-testa kako bismo provjerili odgovaraju li očekivani tempo, očekivano trajanje pjesme i očekivana glasnoća najslušanijih pjesama na Spotifyju u razdoblju 2010. – 2019. aritmetičkim sredinama odgovarajućih vrijednosti za 100 najpopularnijih pjesama na Spotifyju uopće.
Očekivanja i varijance su u svim slučajevima konačne, a set podataka je velik pa su sve pretpostavke Z-testa zadovoljene. Referentne vrijednosti za očekivanja dobili smo kao aritmetičke sredine iz tablice: *Top 100 Most Streamed Songs on Spotify*

Pretpostavke su sljedeće:

H_0: Očekivani tempo najpopularnijih pjesama 2010. – 2019. jednak je aritmetičkoj sredini tempa 100 najpopularnijih pjesama svih vremena.
H_1: Ne vrijedi H_0.

H_0: Očekivano trajanje najpopularnijih pjesama 2010. – 2019. jednako je aritmetičkoj sredini trajanja 100 najpopularnijih pjesama svih vremena.
H_1: Ne vrijedi H_0.

H_0: Očekivana glasnoću najpopularnijih pjesama 2010. – 2019. jednaka je aritmetičkoj sredini glasnoća 100 najpopularnijih pjesama svih vremena.
H_1: Ne vrijedi H_0.

Testnu statistiku računali smo prema formuli

$$Z = \frac{\overline{X_n} - \mu_0}{S_n}\sqrt{n}$$

gdje je $\overline{X_n}$ aritmetička sredina podataka iz naše tablice, μ_0 vrijednost s kojom uspoređujemo očekivanja naših varijabli (u našem slučaju aritmetičke sredine podataka iz referentne tablice), S_n uzoračka standardna devijacija podataka iz naše tablice, a $n = 1000$ broj podataka u našoj tablici.

	tempo	trajnje	glasnoća
pretpostavljeno očekivanje	116.97	214.53	-6.1
srednja vrijednost uzorka	121.262	220.406	-5.663
testna statistika Z	0.19715	0.11656	3.3693
95%-tni pouzdani interval	[119.8969, 122.6271]	[217.2447, 223.5673]	[-5.671133, -5.654867]
p-vrijednost	0.8437	0.9072	0.0007537

Tablica 4.1: Rezultati Z-testova

Na temelju Z-testova u slučaju tempa i trajanja ne odbacujemo H_0, tj. zaključujemo da se očekivani tempo i trajanje najpopularnijih pjesama iz razdoblja 2010. – 2019. poklapaju s aritmetičkom sredinom tempa i trajanja za 100 najslušanijih pjesama svih vremena.
S druge strane, za glasnoću odbacujemo H_0, dakle očekivana glasnoća najpopularnijih pjesama svih vremena se razlikuje od očekivane glasnoće najpopularnijih pjesama 2010.-2019. Kada pogledamo srednju vrijednost našeg i referentnog uzorka, zaključujemo da su najpopularnije pjesme svih vremena nešto tiše.

4.2 χ^2-test homogenosti za frekvenciju žanrova

Zanima nas utječe li godina na trendove određenih žanrova.
Promatramo sljedeće žanrove: dance pop, pop i hip hop. Iz popisa najpopularnijih pjesama određujemo broj pojavljivanja pjesama određenog žanra po svakoj godini. Bitno je napomenuti da se na popisu nalazi po 100 pjesama od svake godine. Zbog toga smo sigurni da ovim testom zaista testiramo trendove. Dobili smo sljedeće podatke (frekvencije):

	Dance pop	Pop	Hip hop
2010.	42	4	1
2011.	45	4	3
2012.	37	4	0
2013.	37	4	1
2014.	39	9	0
2015.	38	5	3
2016.	48	3	1
2017.	25	10	6
2018.	28	3	3
2019.	22	11	3

Tablica 4.2: Broj pojavljivanja žanra po godinama

Možemo li tvrditi da imaju istu razdiobu po svim godinama s pouzdanošću od 5%?

Precizirajmo hipoteze:

H_0: Promatrani žanrovi imaju istu razdiobu.

H_1: Promatrani žanrovi nemaju istu razdiobu.

Test odrađujemo u R-u koristeći χ^2-test o homogenosti.

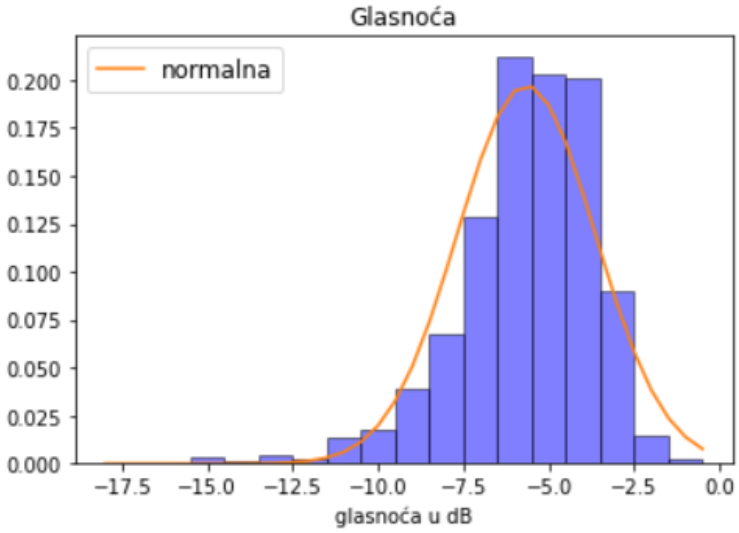
```
> mygenres <- c("dance pop", "pop", "hip hop")
> myyears <- 2010:2019
> years <- split(df, f = df$top.year)
> yearmatrix <- matrix(,nrow=0,ncol=length(mygenres),byrow=TRUE)
> for (i in myyears-2009) {
+   yeardata <- table(years[[i]]$top.genre)
+   yearmatrix <- rbind(yearmatrix, as.vector(yeardata[mygenres]))
+ }
> yearmatrix[is.na(yearmatrix)] <- 0
> yearmatrix
      [,1] [,2] [,3]
 [1,]   42    4    1
 [2,]   45    4    3
 [3,]   37    4    0
 [4,]   37    4    1
 [5,]   39    9    0
 [6,]   38    5    3
 [7,]   48    3    1
 [8,]   25   10    6
 [9,]   28    3    3
[10,]   22   11    3
> chisq.test(yearmatrix)

        Pearson's Chi-squared test

data:  yearmatrix
X-squared = 42.082, df = 18, p-value = 0.001077
```

Dobivamo da je p-vrijednost $0.001077 < 0.05 = 5\%$, stoga odbacujemo hipotezu H_0.

4.3 Kolmogorov – Smirnovljev test pripadnosti glasnoće konkretnoj normalnoj distribuciji

U ovom ćemo odjeljku na razini značajnosti od 5% testirati pripadaju li podaci odgovarajućoj normalnoj razdiobi koristeći Kolmogorov – Smirnovljev test. Kao motivacija poslužio nam je graf sa Slike 4.1.

Neka je $X_1, \ldots, X_{1000}$ slučajni uzorak koji pripada slučajnoj varijabli X koja predstavlja glasnoću. Neka je $x_1, \ldots, x_{1000}$ realizacija slučajnog uzorka u vidu naših podataka. Narančasta linija na grafu prikazuje funkciju gustoće normalno distribuirane slučajne varijable $Y \sim N(\mu, \sigma^2)$, pri čemu su $\mu = \bar{x}$, a $\sigma^2 = \frac{1}{999} S_{XX}$. Navedene parametre koristili smo jer su upravo oni nepristrani procjenitelji očekivanja, odnosno varijance.



Slika 4.1: Usporedba histograma glasnoće i grafa normalne razdiobe s odgovarajućim parametrima

Navedimo hipoteze:

$$H_0\text{: } F = F_0$$
$$H_1\text{: ne } H_0$$

Pritom je F_0 funkcija distribucije normalne slučajne varijable s očekivanjem μ i varijancom σ^2 kao gore. Sam test provodimo u Pythonu, a postupak i rezultate prilažemo na Slici 4.2.

```
mi = np.mean(glasnoca['dB'])
sigma = math.sqrt(np.var(glasnoca['dB']))
stat, pval = stats.ks_1samp(glasnoca['dB'], stats.norm.cdf,[mi,sigma])
print("stat = ",stat," p-vrijednost = ",pval)
```

```
stat =  0.15688779931023034  p-vrijednost =  5.785181065199841e-22
```

Slika 4.2: Kolmogorov – Smirnovljev test za glasnoću

Primijetimo da p-vrijednost iznosi $5.785 \cdot 10^{-22} < 0.05$. Prema tome, odbacujemo hipotezu H_0.

4.4 Lillieforsov test pripadnosti trajanja normalnoj distribuciji

Na temelju opisne statistike primjetili smo da bi trajanja pjesama za svaku godinu mogla biti normalno distribuirane slučajne varijable pa ćemo tu pretpostavku testirati Lillieforsovim

testom pripadnosti nekoj normalnoj distribuciji.

H_0: Trajanje pjesama za svaku od godina 2010. – 2019. normalno je distribuirano.

H_1: Ne vrijedi H_0

godina	p-vrijednost	p-vrijednost > 0.01
2010.	0.001282	$-$
2011.	0.03022	$+$
2012.	0.0005963	$-$
2013.	$1.962e-05$	$-$
2014.	$1.79e-07$	$-$
2015.	0.6432	$+$
2016.	0.03877	$+$
2017.	0.01688	$+$
2018.	0.0002701	$-$
2019.	0.0001166	$-$

Tablica 4.3: Rezultati Lillieforsovih testova

Ako uzmemo razinu značajnosti $\alpha = 0.01$, za četiri godine ne možemo odbaciti pretpostavku da se radi o nekoj normalnoj distribuciji. S ovakvim rezultatom nismo previše zadovoljni, ali kako su se, od svih podataka kojima baratamo, ovi pokazali najbliži normalnoj distribuciji, zaključit ćemo da za godine 2011., 2015., 2016. i 2017. ne možemo na statistički značajnoj razini odbaciti hipotezu H_0 o pripadnosti trajanja normalnoj distribuciji.

4.5 F-test usporedbe varijanci trajanja pjesama za odabrane godine

Nastavit ćemo s usporedbom varijanci trajanja pjesama za četiri godine za koje nismo isključili pripadnost normalnoj distribuciji. Provest ćemo ukupno šest F-testova kako bismo testirali jednakost svakog para varijanci.

U svakom testu hipoteze su:

$$H_0\colon \sigma_1 = \sigma_2$$
$$H_1\colon \sigma_1 \neq \sigma_2$$

godine	F	$\sigma_1 : \sigma_2$	95%-tni pouzdani interval	p
2011./2015.	1.5845	1.584485	[1.066107, 2.354915]	0.02296
2011./2016.	1.1365	1.136509	[0.7646904, 1.6891177]	0.5256
2011./2017.	0.73622	0.73622	[0.4953595, 1.0941951]	0.1294
2015./2016.	0.71727	0.7172734	[0.4826114, 1.0660359]	0.09992
2015./2017	0.46464	0.4646432	[0.3126313, 0.6905684].	0.0001708
2016./2017.	0.64779	0.6477909	[0.4358607, 0.9627687]	0.03185

Tablica 4.4: Rezultati F-testova

p-vrijednosti koje se ne odnose na 2015. godinu su sve relativno velike, stoga zaključujemo da na razini značajnosti od 97% ne možemo odbaciti pretpostavku jednakosti varijanci u skupini {2011., 2016., 2017.}. Neke p-vrijednosti koje uključuju 2015. godinu su velike (0.09992), a neke male (< 0.0002). Nama su zanimljivije ove manje jer nam omogućuju da odbacimo prepostavku o jednakosti varijanci, konkretno za 2015. i 2017. godinu. Krajnji zaključak je da se varijance trajanja pjesama za godine 2011., 2016. i 2017. ne razlikuju statistički značajno, dok varijanca godine 2015. odudara od njih.

4.6 ANOVA test za usporedbu očekivanog trajanja pjesama za 2011., 2016. i 2017.

Rezultate o normalnoj distribuciji i jednakosti varijanci iz prethodnih dvaju poglavlja koristimo kako bismo opravdali provedbu ANOVA testa za usporedbu očekivanog trajanja pjesama za 2011., 2016. i 2017. godinu. Hipoteze želimo interpretirati na nivou značajnosti $\alpha = 0.05$.

H_0: Očekivana trajanja pjesama za 2011., 2016. i 2017. godinu su jednaka.

H_1: Postoje barem dvije godine čija se očekivana trajanja pjesama razlikuju.

```
Analysis of Variance Table

Response: trajanje
           Df Sum Sq Mean Sq F value Pr(>F)
grupe       2   2470  1234.8  1.0844 0.3394
Residuals 297 338203  1138.7
```

Slika 4.3: Rezultati provedenog ANOVA testa

Testna F-statistika iznosi $f = F(k, n-k) = F(2,298) = 1.0844$, dok iz tablica pročitamo vrijednost $f_{0.05}(2,298) = 4.61$. Kako je $f = 1.0844 < 4.61 = f_{0.05}(2,298)$, f ne upada u kritično područje, stoga na razini značajnosti $\alpha = 0.05$ ne odbacujemo hipotezu H_0.

4.7 Lillieforsov test pripadnosti tempa normalnoj distribuciji

Provjerimo Lillieforsovim testom jesu li tempa pjesama po godinama normalno distibuirane slučajne varijable. Testiramo sljedeće hipoteze na razini značajnosti 0.05.

H_0: Tempo pjesama za svaku od godina 2010.-2019. je normalno distibuiran.

H_1: Ne vrijedi H_0

godina	p-vrijednost	p-vrijednost > 0.01
2010.	0.001033	$-$
2011.	0.0001376	$-$
2012.	$8.736e-06$	$-$
2013.	0.0003998	$-$
2014.	$2.407e-05$	$-$
2015.	0.0037	$-$
2016.	0.0001937	$-$
2017.	0.05699	$+$
2018.	0.1435	$+$
2019.	0.002588	$-$

Tablica 4.5: Rezultati Lillieforsovih testova

Iz tablice se vidi kako je p-vrijednost za podatke 2017. i 2018. godine veća od razine značajnosti $\alpha = 0.05$ zbog čega samo za navedene godine možemo zaključiti kako ne možemo odbaciti hipotezu H_0 o pripadnosti normalnoj distribuciji.

4.8 T-test za usporedbu očekivanja tempa pjesma iz 2017. i 2018. godine

Pretpostavljamo da su tempa pjesama iz 2017. i 2018. godine normalno distibuirana po diskusiji iz odjeljka 4.7 te da imaju jednake varijance. T-testom testiramo očekivanja tempa

zadanih godina, pri čemu X_1 predstavlja tempa pjesama iz 2017., $\mu_1 = \mathbb{E}X_1$, X_2 tempa pjesama iz 2018., $\mu_2 = \mathbb{E}X_2$.
Hipoteze su sljedeće:

$$H_0\colon \mu_1 = \mu_2$$
$$H_1\colon \mu_1 \neq \mu_2$$

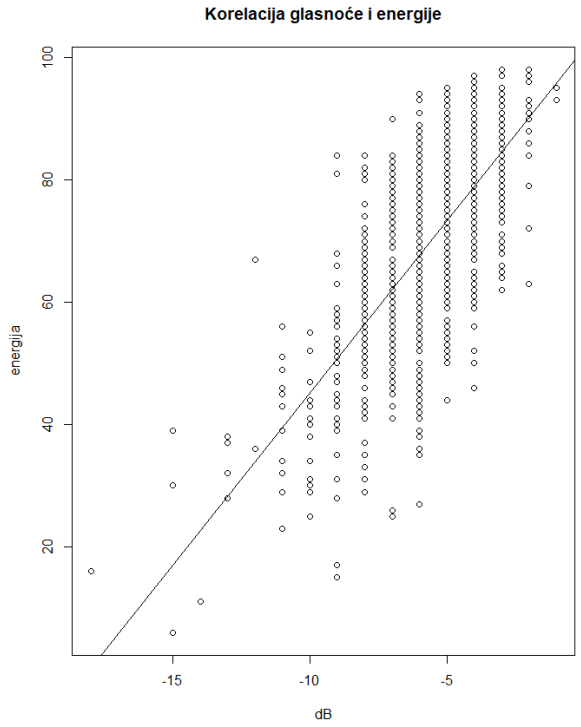
```
        Two Sample t-test

data:  tempo8$`125` and tempo9$`102`
t = 0.38086, df = 198, p-value = 0.7037
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -6.141334  9.081334
sample estimates:
mean of x mean of y
   122.02    120.55
```

Slika 4.4: Rezultat T-testa tempa

Kako je p-vrijednost $= 0.7037 > 0.05$, ne odbacujemo hipotezu H_0 u korist hipoteze H_1.

4.9 Linearna regresija za glasnoću i energiju pjesama

Promotrimo sada promjenu energetičnosti u odnosu na promjenu glasnoće. Pearsonov koeficijent korelacije iznosi 0.7134284, tako da ima smisla raditi linearnu regresiju.



Slika 4.5: Linearna interpolacija energije i glasnoće

Pravac koji najbolje aproksimira podatke zadan je sa $y = \hat{\alpha} + \hat{\beta}x$ za $\hat{\alpha} = 101.343628$ i $\hat{\beta} = 5.622749$.

Linearnu regresiju odrađujemo u R-u:

```
> regr=lm(df$energy~df$dB)
> plot(df$dB, df$energy, xlab = "dB", ylab = "energija")
> abline(regr$coefficients)
> plot(df$dB, df$energy, xlab = "dB", ylab = "energija", main="Korelacija glasnoće i energije")
> abline(regr$coefficients)
> print(summary(regr))

Call:
lm(formula = df$energy ~ df$dB)

Residuals:
    Min      1Q  Median      3Q     Max
-40.607  -7.230   0.393   7.770  33.261

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 101.3436     1.0513   96.40   <2e-16 ***
df$dB         5.6227     0.1748   32.16   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.19 on 998 degrees of freedom
  (3 observations deleted due to missingness)
Multiple R-squared:  0.509,     Adjusted R-squared:  0.5085
F-statistic:  1035 on 1 and 998 DF,  p-value: < 2.2e-16

> confint(regr)
                2.5 %     97.5 %
(Intercept) 99.280558 103.406697
df$dB        5.279699   5.965799
```

Odredimo sada pouzdane intervale za vrijednosti

	fit	lwr	upr
-15	17.00239	13.72493	20.27986
-14	22.62514	19.68204	25.56824
-13	28.24789	25.63690	30.85888
-12	33.87064	31.58852	36.15276
-11	39.49339	37.53527	41.45151
-10	45.11614	43.47426	46.75802
-9	50.73889	49.39998	52.07779
-8	56.36164	55.30101	57.42227
-7	61.98438	61.15218	62.81659
-6	67.60713	66.90317	68.31110
-5	73.22988	72.49918	73.96059
-4	78.85263	77.95393	79.75133
-3	84.47538	83.32788	85.62288
-2	90.09813	88.66243	91.53383
-1	95.72088	93.97702	97.46474

Slika 4.6: Pouzdani intervali

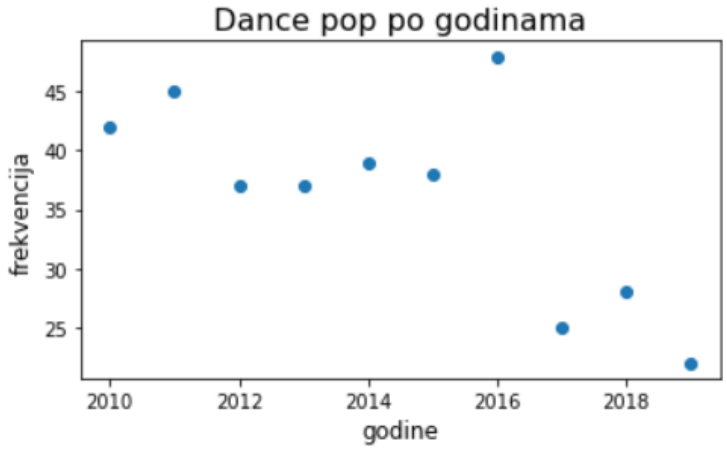
4.10 Linearna regresija za udio dance pop žanra po godinama

Promotrimo sada promjenu broja pjesama najzastupljenijeg žanra - dance popa - u vremenu. Na Slici 4.7 dana je tablica frekvencija dance popa po godinama.

godine	2010.	2011.	2012.	2013.	2014.	2015.	2016.	2017.	2018.	2019.
frekvencija	42	45	37	37	39	38	48	25	28	22

Slika 4.7: Zastupljenost dance pop žanra po godinama

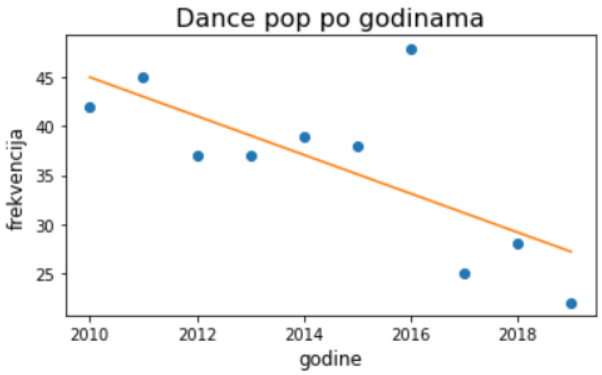
Iz tablice možemo iščitati da se s godinama popularnost dance pop žanra smanjuje. Nacrtajmo graf da vidimo postoji li pravilnost.



Slika 4.8: Graf dance pop žanra po godinama

Pearsonov koeficijent korelacije iznosi -0.70232989, stoga ima smisla raditi linearnu regresiju.

```
n = np.size(x)
x_s = np.mean(x)
y_s = np.mean(y)
Sxx = np.sum(x*x) - n*x_s*x_s
Syy = np.sum(y*y) - n*y_s*y_s
Sxy = np.sum(x*y) - n*x_s*y_s

beta = Sxy/Sxx
alfa = y_s - beta*x_s
plt.plot(x, beta*x + alfa)
```



Slika 4.9: Linearna regresija

Pravac koji najbolje aproksimira podatke je $y = \hat{\alpha} + \hat{\beta}x$ za $\hat{\alpha} = 4028.473$ i $\hat{\beta} = -1.982$.
Postupkom kao na Slici 4.10 dobivamo da su 95% pouzdani intervali za $\hat{\alpha}$ i $\hat{\beta}$ redom $[729.377, 7327.569]$ i $[-3.619, -0.344]$. Možemo primijetiti da su intervali prilično široki. Razlog tome je što koeficijent korelacije između godina i dance pop žanra, kao i sam uzorak nisu preveliki.

```
SSE = Syy - beta*beta*Sxx
sigma = math.sqrt(SSE/(n-2))
vrij = stats.t.ppf(0.975,n-2)
naz_a = math.sqrt(1/n + x_s*x_s/Sxx)
naz_b = math.sqrt(1/Sxx)

poc_a = alfa - vrij * sigma * naz_a
kraj_a = alfa + vrij * sigma * naz_a
poc_b = beta - vrij * sigma * naz_b
kraj_b = beta + vrij * sigma * naz_b
```

Slika 4.10: Račun intervala pouzdanosti za parametre regresije

Alternativno, mogli smo godine translatirati tako da podaci na x-osi budu $0, 1, \ldots, 9$. Koeficijent $\hat{\beta}$ ne mijenja se jer svakako moramo dobiti jednaki pravac (točnije, aditivna konstanta u računu uzoračke varijance i kovarijance ne utječe na konačni rezultat), no koeficijent $\hat{\alpha}$, kao i njegov interval pouzdanosti, promijenit će se, s obzirom da ga računamo po formuli $\hat{\alpha} = y_m - \hat{\beta} * (x_m - 2010)$
Novi procjenitelj tako iznosi $\hat{\alpha} = 45.018$, a interval pouzdanosti $[36.275, 53.761]$.

5 Zaključak

Za početak prokomentirajmo mogu li se naši rezultati generalizirati na cijelu populaciju. Spotify je 2015. imao 18 milijuna korisnika diljem svijeta te je bio u porastu. U odnosu na tadašnjih 7.3 milijardi ljudi na svijetu, to predstavlja čak $0,25\%$ svjetske populacije. Danas je taj udio jednak $2,25\%$ (tj. 180 milijuna korisnika od 8 milijardi ljudi na svijetu).
Dakle, rezultati se zaista mogu generalizirati, ali ne i potpuno. Naime, Spotify usluga se plaća te možemo pretpostaviti da si velik dio svjetske populacije neće moći priuštiti takvu uslugu. Također, Spotify je aplikacija izrađena u Stockholmu 2006., stoga možemo naslutiti (ovu tvrdnju nećemo testirati) da je prvobitna publika na Spotifyju bila europska te vrlo brzo i američka, a kasnije se tek širila dalje po svijetu. Promjena publike svakako ima utjecaja na najslušanije pjesme te jedini razlog nije nužno samo promjena trenda u cijeloj populaciji.
Mi pretpostavljamo da je na podatke grupirane po godinama zaista utjecalo i generalno mijenjanje trendova jer je broj pretplatnika na Spotify zbilja velik.

Prokomentirajmo sada naše zaključke.
Pomoću χ^2 - testa o homogenosti prvo smo zaključili da se trendovi žanrova značajno mijenjaju po godinama te da zastupljenost pjesama po žanrovima nije rezultat varijance podataka. Zatim smo opovrgnuli da je glasnoća pjesama normalno distribuirana varijabla. Normalnu distribuiranost tempa i trajanja pjesama za neke smo godine opovrgnuli, dok za druge to nismo uspjeli. Za godine kod kojih nismo opovorgnuli normalnost pokušali smo opovrgnuti jednakost varijanci pomoću F-testa. Naposljetku smo za tri godine za koje nismo uspjeli opovrgnuti jednakost varijanci pokušali ANOVA-om pokazati da postoji godina čije se očekivano trajanje pjesma razlikuje od očekivanog trajanja preostalih. Nismo uspjeli pokazati da postoji takva godina na statistički značajnoj razini. T-testom hipotezu da su očekivanja tempa dvije godine jednaka nismo odbacili u korist alternative.
Pomoću Z-testova usporedbe očekivanja zaključili smo da se očekivani tempo i trajanje pjesama iz razdoblja $2010. - 2019.$ ne razlikuju statistički značajno od aritmetičke sredine podataka najslušanijih pjesama svih vremena. Dakle, određene karakteristike pjesama ipak ostaju iste kroz sve godine. S druge strane, zaključili smo da se glasnoća pjesama koje su zadnjih godina popularne razlikuje od svevremenskih hitova.
Na kraju smo vidjeli da su energija pjesme i glasnoća pozitivno linearno korelirani pa pretpostavljamo da glasnoća utječe na naš dojam o tome koliko je pjesma energična.
Također, linearna je regresija pokazala da zastupljenost dance pop žanra na top ljestvicama s godinama opada, iako je još uvijek među svim žanrovima dominantan.